2022 AP Statistics Exam Questions

Extrated from the AP Question Bank

Contents

1 Multiple Choice Questions

The exam begins on the next page.

STATISTICS

SECTION I

Time—1 hour and 30 minutes

40 Questions

Directions: Each of the questions or incomplete statements below is followed by five suggested answers or completions. Select the one that is best in each case and then fill in the corresponding circle on the answer sheet.

1. A university that offers evening classes to a large number of students who commute to campus wants to estimate the mean commute time of all evening students. A parking facilities attendant selects one parking lot on campus, then selects the first 60 students who drive into the lot on a certain evening. Each of the 60 students is asked to provide their typical commute time to campus. The university will create a confidence interval for the mean commute time of evening students.

   Are the conditions for inference satisfied?

   (A) Yes, all conditions for inference are satisfied.
   (B) No, because the distribution of the population mean commute time is not known to be normally distributed.
   (C) No, because the sample size is greater than 10% of the population size.
   (D) No, because the students selected are not a random sample of all evening students.
   (E) No, because the sample size of 60 is not large enough to assume that the sampling distribution of the sample mean commute time is approximately normal.

2. The distribution of the length of employment for workers at a factory is moderately skewed to the right, with mean 7.4 years and standard deviation 5.6 years. The factory manager will select a random sample of 100 employees to survey.

   Which of the following best describes the sampling distribution of the sample means of length of employment for samples of size 100?

   (A) Moderately skewed to the right, with mean 7.4 years and standard deviation 5.6 years
   (B) Moderately skewed to the right, with mean 7.4 years and standard deviation less than 5.6 years
   (C) Moderately skewed to the right, with mean less than 7.4 years and standard deviation less than 5.6 years
   (D) Approximately normal, with mean less than 7.4 years and standard deviation 5.6 years
   (E) Approximately normal, with mean 7.4 years and standard deviation less than 5.6 years

3. Researchers from a national research organization investigated which platform (television, social media, newspaper, Internet, etc.) United States adults prefer to use to obtain their news. In 2016, of the 200 randomly selected United States adults that were surveyed, 114 selected television as their preferred platform. In 2018, of the 200 randomly selected United States adults that were surveyed, 98 selected television as their preferred platform. The conditions for inference are met. A 95 percent confidence interval for the difference in the population proportions of United States adults that use television as their preferred platform for obtaining news between 2016 and 2018 (2016 minus 2018) was found to be $(-0.02, 0.18)$.

Based on the interval, is it reasonable for the researchers to claim that there is convincing statistical evidence that the population proportion of United States adults that use television as their preferred method to obtain news differed between 2016 and 2018?

(A) No, because the researchers should have used a greater confidence level to obtain convincing evidence.

(B) No, because zero is included in the interval there is not convincing statistical evidence that the population proportion differed between 2016 and 2018.

(C) Yes, because zero is included in the interval there is convincing statistical evidence that the population proportion differed between 2016 and 2018.

(D) Yes, because most of the interval is positive there is convincing statistical evidence that the population proportion increased from 2016 to 2018.

(E) Yes, because 0.57 is greater than 0.49 there is convincing statistical evidence that the population proportion decreased from 2016 to 2018.

4. A solar power engineer took a random sample of houses and installed the same type of solar panels using two different methods on each house to investigate whether there is a mean difference in the angles of installation between the two methods for all houses in the population of interest. The engineer found the sample mean difference between the two methods to be 0.2 degree and the p-value for a two-sided matched-pairs t-test for the mean difference to be 0.65.

Assuming the conditions for inference are met, which of the following statements is the best interpretation of the p-value?

(A) The probability that the null hypothesis is true is 0.65.

(B) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of 0.2 degree or -0.2 degree.

(C) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of greater than 0.2 degree or less than -0.2 degree.

(D) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of greater than 0.2 degree.

(E) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of less than -0.2 degree.

5. Each accountant at a large accounting firm was classified according to accountant level (junior or senior) and method of transportation to work (walk, bus, drive alone, or carpool). The responses of the 320 accountants at the firm are summarized in the table.

	Junior	**Senior**	**Total**
Walk	25	3	28
Bus	87	12	99
Drive alone	96	25	121
Carpool	52	20	72
Total	260	60	320

What proportion of the accountants at the firm are at the senior level and carpool to work?

(A) $\frac{20}{60}$

(B) $\frac{20}{72}$

(C) $\frac{20}{320}$

(D) $\frac{112}{320}$

(E) $\frac{132}{320}$

6. The following table shows the classification of all 51 dogs from an animal shelter by whether the dogs lived mostly outdoors or mostly indoors before coming to the shelter and whether they tested positive or negative for a certain skin condition.

	Positive	**Negative**	**Total**
Mostly Outdoors	10	20	30
Mostly Indoors	8	13	21
Total	18	33	51

One dog from the animal shelter will be selected at random. Based on the table, which of the following statements is correct?

(A) The probability of selecting a dog that tested positive is greater than 0.5.

(B) The probability of selecting a dog that lived mostly indoors is greater than 0.5.

(C) The probability of selecting a dog that tested positive is greater than the probability of selecting a dog that tested negative.

(D) The probability of selecting a dog that tested positive given that the dog lived mostly outdoors is less than the probability of selecting a dog that tested positive given that the dog lived mostly indoors.

(E) The probability of selecting a dog that lived mostly outdoors given that the dog tested positive is greater than the probability of selecting a dog that lived mostly outdoors given that the dog tested negative.

7. Researchers are examining the relationship between hours of sleep and athletic performance among college athletes. Athletic performance will be measured on a numeric scale, with greater numbers indicating better performance. The researchers expect that the more hours the athletes sleep, the better they will perform. Assuming all conditions for inference are met, the researchers will create a 95 percent confidence interval for the slope of the regression line for predicting athletic performance from amount of sleep.

   For which of the following would the confidence interval support the researchers' expectations?

   (A) The confidence interval includes only positive values.
   (B) The confidence interval includes only negative values.
   (C) The confidence interval has a width less than 1.
   (D) The confidence interval has a width greater than 1.
   (E) The confidence interval includes the value 0.

8. A company that sells baby food is interested in the baby food preferences of all families with toddlers from a certain city. A representative from the company sets up a booth at one grocery store in the city that will be used to investigate baby food preference. The representative will ask people with toddlers who walk past the booth to complete the survey.

   Which of the following best describes the sampling method the company will use?

   (A) Systematic random sampling
   (B) Simple random sampling
   (C) Cluster sampling
   (D) Stratified random sampling
   (E) Convenience sampling

9. A large online retailer places packed and ready-to-be-shipped boxes in a line on a conveyor belt. A sample of boxes will be used to estimate the mean weight of boxes shipped on a given day. One proposed method for selecting the sample is as follows:

   - Generate a random integer from 1 to 20.
   - Use that integer to select an initial box from the first 20 boxes in line on the conveyor belt.
   - Select every 25th box that follows the initial box on the conveyor belt for the rest of the day.

   Which of the following best describes the proposed sampling method?

   (A) Stratified random sample, stratified by time of day
   (B) Cluster sample, with times of day as clusters
   (C) Cluster sample, with box weights as clusters
   (D) Systematic random sample
   (E) Convenience sample

10. Each voter from a random sample of 334 registered voters was asked their impression of two candidates running for the same national office. The table summarizes the responses.

	Candidate A	**Candidate B**
Favorable	138	200
Unfavorable	44	47
No opinion	88	47
Have not heard of	64	40

Which of the following is the most appropriate method to use to estimate the proportion of all registered voters who have a favorable impression of Candidate A?

(A) A one-sample z-interval for estimating a sample proportion

(B) A one-sample z-interval for estimating a population proportion

(C) A matched-pairs t-interval for estimating a mean difference

(D) A two-sample z-interval for estimating a difference between sample proportions

(E) A two-sample z-interval for estimating a difference between population proportions

11. A raffle for a charity fund-raiser is being planned. Each of 2000 raffle tickets will be sold for \$1.00. The holders of 32 winning tickets will each win a prize. The table shows the prize values and the number of prizes for each value.

Prize Value	**Number of Prizes**
\$25	20
\$50	10
\$300	2

The random variable W represents the value of the prize won for a single ticket minus the cost of the ticket. What is the expected value of W?

(A) $-\$1.00$

(B) $-\$0.20$

(C) \$0.80

(D) \$49.00

(E) \$124.00

12. In a study researching how donating to charity can affect a person's happiness, 96 participants were given $5 a day for one week. Each participant was randomly assigned to one of two groups. Those assigned to the first group were asked to spend the money on themselves, and those assigned to the second group were asked to donate the money to charity. At the end of the week, all of the participants were asked to rate their overall level of happiness on a scale from 0 to 100, with higher scores indicating greater levels of happiness.

    Which of the following are the appropriate null and alternative hypotheses to test whether the sample data provide convincing statistical evidence that donating to charity results in greater happiness than spending on oneself, on average?

    (A) H_0: $\mu_{\text{self}} - \mu_{\text{charity}} = 0$
    H_a: $\mu_{\text{self}} - \mu_{\text{charity}} > 0$

    (B) H_0: $\mu_{\text{self}} - \mu_{\text{charity}} = 0$
    H_a: $\mu_{\text{self}} - \mu_{\text{charity}} < 0$

    (C) H_0: $\mu_{\text{self}} - \mu_{\text{charity}} = 0$
    H_a: $\mu_{\text{self}} - \mu_{\text{charity}} \neq 0$

    (D) H_0: $\mu_{\text{self}} - \mu_{\text{charity}} < 0$
    H_a: $\mu_{\text{self}} - \mu_{\text{charity}} = 0$

    (E) H_0: $\mu_{\text{self}} - \mu_{\text{charity}} > 0$
    H_a: $\mu_{\text{self}} - \mu_{\text{charity}} < 0$

13. Suppose a hypothesis test will be used to investigate whether the proportion of United States adults who have been a victim of a cybercrime, such as identity theft, is greater than 0.25. Suppose that the hypothesis test is conducted using a significance level of 0.05 and the power of the test is determined for a specific alternative value using the significance level of 0.05.

    If the significance level of the hypothesis test is changed to 0.01 and the power of the test is computed for the same alternative value using a significance level of 0.01, which of the following best describes the change(s) to the probability of a Type I error and the power of the test?

    (A) The probability of a Type I error would decrease, and the power of the test would stay the same.

    (B) The probability of a Type I error would decrease, and the power of the test would increase.

    (C) The probability of a Type I error would decrease, and the power of the test would decrease.

    (D) The probability of a Type I error would increase, and the power of the test would increase.

    (E) The probability of a Type I error would increase, and the power of the test would decrease.

14. A large company has operated assuming that the rate of absenteeism is the same for each day of the week. The new human resources director at the company claims that the assumption is false and that the rate of absenteeism is different for some days of the week as compared with others. To test the claim, a random sample of 150 absences was selected, and the day of the week on which the absence occurred was recorded. The counts are shown in the following table.

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	40	27	21	34	28

The appropriate chi-square test will be performed to test the claim.

What is the contribution of the Monday absences to the calculation of the chi-square test statistic?

(A) 0.333
(B) 2.500
(C) 3.000
(D) 3.333
(E) 10.000

15. The manager of a restaurant plans to survey a random sample of customers to estimate the proportion of customers who would return within a week if given a coupon for a free dessert to use on their next visit. A similar coupon was given last year, and 30 percent of the customers who were given the coupon returned within a week. The manager will use the observed proportion from last year as a preliminary estimate of the population proportion and wants the margin of error to be at most 0.04 at a 90 percent level of confidence.

Of the following, which is the least number of customers who should be surveyed such that the estimate of the margin of error of the 90 percent confidence interval will be no greater than 0.04?

(A) 15
(B) 260
(C) 310
(D) 370
(E) 610

16. An experiment was conducted to investigate whether submersion in cold water causes a lower heart rate. The experiment used 50 volunteers. The 25 youngest volunteers had their heart rate measured while holding their breath for 30 seconds with their face submerged in cold water. The 25 oldest volunteers had their heart rate measured while holding their breath for 30 seconds with their face not submerged in cold water. The mean heart rate for volunteers who had their face submerged in cold water was lower than the mean heart rate for volunteers who did not have their face submerged in cold water.

Which of the following elements of a well-designed experiment is missing?

I. Comparison of at least two treatment groups

II. Random assignment of treatments to experimental units

III. Replication

(A) I only

(B) II only

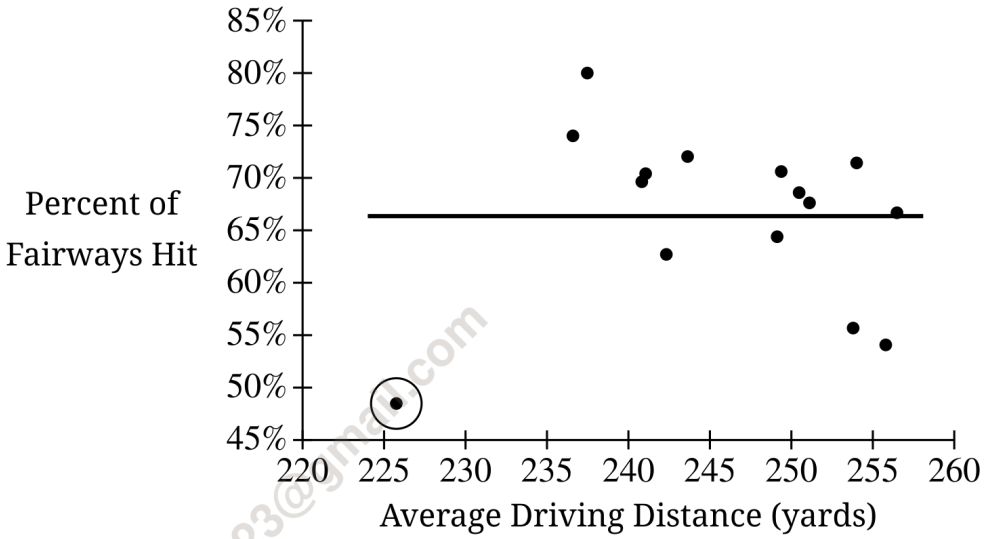
(C) I and II only

(D) II and III only

(E) I, II, and III

17. A sample of 15 golfers who played a golf course on a certain day was selected. For each golfer, the average driving distance (x), in yards, and the percent of fairways hit on the drive (y) were recorded.

The scatterplot displays the percent of fairways hit versus the average driving distance. Also shown is the least-squares regression line, $\hat{y} = 66.228 + 0.0002x$.



The point circled on the scatterplot is considered an influential point. A new least-squares regression line will be calculated with the influential point removed. How will the removal of the influential point affect the new least-squares regression line for the remaining 14 points?

(A) The y-intercept will remain the same, and the slope will be negative.

(B) The y-intercept will decrease, and the slope will be negative.

(C) The y-intercept will decrease, and the slope will be positive.

(D) The y-intercept will increase, and the slope will be negative.

(E) The y-intercept will increase, and the slope will be positive.

18. For the population of people in the United States who are at least 18 years of age, approximately 61 percent voted in a recent election. A random sample of 200 people and a random sample of 100 people will be selected from the population.

For which sample size is it more likely that greater than 70 percent of the sample voted in the recent election?

(A) They are equally likely, because the mean of the sampling distribution is the same for both sample sizes.

(B) It is more likely for a sample size of 200 people, because there is more variability in the sampling distribution for larger sample sizes.

(C) It is more likely for a sample size of 200 people, because there is less variability in the sampling distribution for larger sample sizes.

(D) It is more likely for a sample of size 100 people, because there is more variability in the sampling distribution for smaller sample sizes.

(E) It is more likely for a sample size of 100 people, because there is less variability in the sampling distribution for smaller sample sizes.

19. The probability that a particular electrical component operates successfully is $\frac{3}{4}$. An electrical system consists of three such components that operate independently. The system will function successfully if at least one of the three components operates successfully.

What is the probability that the system will function successfully?

(A) $\frac{1}{64}$

(B) $\frac{27}{64}$

(C) $\frac{37}{64}$

(D) $\frac{3}{4}$

(E) $\frac{63}{64}$

20. A food scientist conducting a study to investigate the nutritional habits of teenagers in a city selected a random sample of 100 teenagers in the city. Each of the 100 teenagers was asked to complete a survey with questions about age and the number of ounces of carbonated drinks consumed on the previous day.

Assuming the conditions for inference have been met, which of the following is an appropriate test and alternative hypothesis to use to investigate whether the average number of ounces of carbonated drinks consumed in a day increases with age for teenagers in the city?

(A) A t-test for the slope of a regression model with alternative hypothesis: H_a: $\beta > 0$

(B) A t-test for the slope of a regression model with alternative hypothesis: H_a: $\beta < 0$

(C) A t-test for the slope of a regression model with alternative hypothesis: H_a: $\beta = 0$

(D) A one-sample z-test for a population proportion with alternative hypothesis: H_a: $p > 0.5$

(E) A two-sample t-test for a difference of two population means with alternative hypothesis: H_a: $\mu_{\text{older}} \neq \mu_{\text{younger}}$

21. For a random sample of 50 newborn Holstein calves from the United States, the mean weight was 80 pounds and the standard deviation was 6 pounds. In order to compare these findings with those of other countries, an agricultural engineer converted the measured calf weights from pounds to kilograms by multiplying each weight by 0.454.

What are the mean and standard deviation of the converted weights of the 50 newborn Holstein calves?

(A) The mean weight is 36.32 kilograms and the standard deviation is 2.724 kilograms.

(B) The mean weight is 36.32 kilograms and the standard deviation is 6 kilograms.

(C) The mean weight is 80 kilograms and the standard deviation is 2.724 kilograms.

(D) The mean weight is 80 kilograms and the standard deviation is 6 kilograms.

(E) The mean weight is approximately 176.21 kilograms and the standard deviation is 6 kilograms.

22. The distribution of housing prices in a city includes several outliers at the upper end of the distribution and no outliers at the lower end of the distribution.

    Which of the following is the most resistant, or robust, measure of center of the distribution of housing prices?

    (A) The mean, because it is affected by the outliers.
    (B) The mean, because it is not greatly affected by the outliers.
    (C) The median, because it is affected by the outliers.
    (D) The median, because it is not greatly affected by the outliers.
    (E) The mode, because the data are categorical.

23. A study reported that 14 percent of cars registered in 2015 had a manual transmission. An automotive researcher collected data on cars registered in 2021 to investigate whether the proportion of registered cars with a manual transmission is less than 0.14. The researcher took a random sample of 150 cars registered in 2021 and found 18 of the cars had a manual transmission.

    After ensuring the conditions for inference were met, the researcher conducted an appropriate hypothesis test and calculated a p-value of 0.24. At the 5 percent level of significance, which of the following conclusions should be made?

    (A) The null hypothesis should be rejected. There is significant evidence that the population proportion of registered cars with a manual transmission in 2021 is less than 0.14.
    (B) The null hypothesis should be rejected. There is not significant evidence that the population proportion of registered cars with a manual transmission in 2021 is less than 0.14.
    (C) The null hypothesis should not be rejected. There is significant evidence that the population proportion of registered cars with a manual transmission in 2021 is not less than 0.14.
    (D) The null hypothesis should not be rejected. There is significant evidence that the population proportion of registered cars with a manual transmission in 2021 is equal to 0.14.
    (E) The null hypothesis should not be rejected. There is not significant evidence that the population proportion of registered cars with a manual transmission in 2021 is less than 0.14.

24. For which of the following distributions is the proportion of values between 900 and 1,000 the greatest?

    (A) A normal distribution with mean 900 and standard deviation 100
    (B) A normal distribution with mean 950 and standard deviation 100
    (C) A normal distribution with mean 800 and standard deviation 100
    (D) A normal distribution with mean 1,000 and standard deviation 400
    (E) A normal distribution with mean 950 and standard deviation 50

25. An owner of a novelty business operates both an online store and a retail store. The owner took a random sample of 50 sales from each store during the past year and recorded the purchase amount of each sale. Summary statistics are shown in the following table.

	Online	**Retail**
n	50	50
Mean	\$42.80	\$27.50
Standard deviation	\$10.50	\$8.25

Assuming the conditions for inference were met, the owner calculated a 95 percent confidence interval for the difference in mean purchase amounts (online minus retail) as $(\$11.55, \$19.05)$.

Which of the following is an appropriate interpretation of the 95 percent confidence interval?

(A) We are 95% confident that the purchase amount from a randomly selected online store sale will be between \$11.55 and \$19.05 greater than the purchase amount from a randomly selected retail store sale.

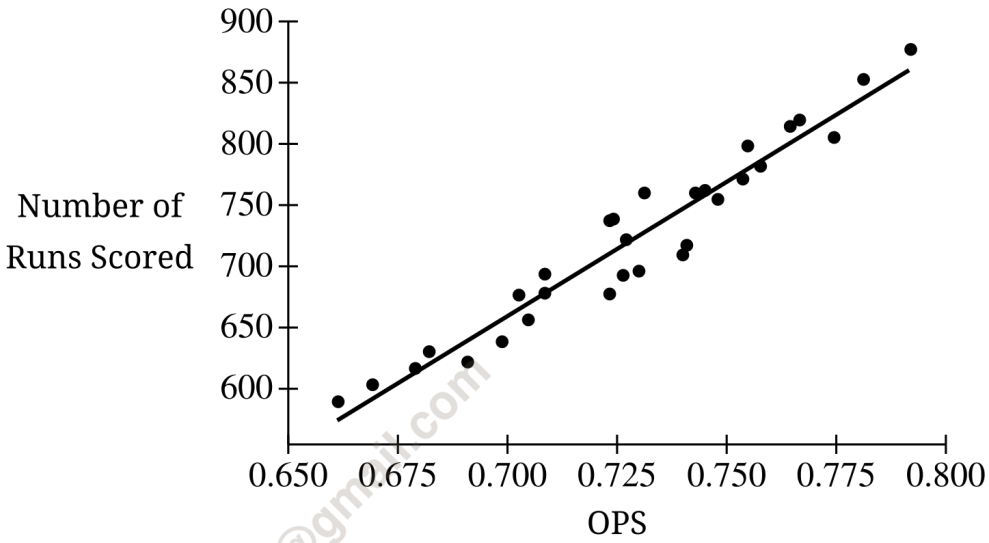
(B) We are 95% confident that the sample mean purchase amount from the online store is between \$11.55 and \$19.05 greater than the sample mean purchase amount from the retail store for the past year.

(C) We are 95% confident that the population mean purchase amount from the online store is between \$11.55 and \$19.05 greater than the population mean purchase amount from the retail store for the past year.

(D) We are 95% confident that the population mean purchase amount for the past year from both stores combined is between \$11.55 and \$19.05.

(E) We are 95% confident that if new samples of the same size were taken from each store, the sample mean purchase amount from the online store would be between \$11.55 and \$19.05 greater than the sample mean purchase amount from the retail store.

26. On-base percentage plus slugging (OPS) is a statistic used in baseball to measure a team's batting success. The number of runs scored and OPS for 30 baseball teams was used to conduct a linear regression analysis. The scatterplot and computer output for the regression analysis is shown below.



Term	Coef	SE Coef	T-Value	P-Value
Constant	−838.40	77.99	−10.75	<0.0001
OPS	2144.43	107.1	20.01	<0.0001
S= 19.516		R-Sq=93.47%		R-Sq(adj)= 93.23%

Which of the following is the most appropriate interpretation of the statistic 93.47% in the regression output?

(A) There is a strong, positive, linear relationship between number of runs scored and OPS.

(B) The typical deviation between observed and predicted number of runs scored is 0.9347.

(C) For each one-unit increase in OPS, the regression model predicts an increase of 93.47 runs scored.

(D) 93.47% of the observed number of runs scored are close to the regression line.

(E) 93.47% of the variation in number of runs scored can be explained by the linear regression with OPS.

27. A paint manufacturer conducted an experiment to investigate whether a new formula produces paint that lasts longer on the exterior of a building than paint produced using an older formula. A grid of 20 squares was drawn on one side of an exterior wall of a building. The new formula was randomly assigned to 10 squares, and the older formula was assigned to the remaining 10 squares. Each square was painted with the assigned formula. One year later, each square was rated on how well the paint had lasted, using a numerical scale for which larger values indicated longer-lasting paint.

Which of the following best describes the design of the experiment?

(A) A completely randomized design, because the two formulas were compared.

(B) A completely randomized design, because each formula was randomly assigned to 10 squares.

(C) A matched pairs design, because the same number of squares were assigned to each formula.

(D) A matched pairs design, because two formulas were compared.

(E) A randomized block design, because the wall was partitioned into 20 squares.

28. A 95 percent confidence interval for the proportion of parents who use parental controls for blocking, filtering, or monitoring their teenagers' online activities is (0.36,0.42).

Which of the following could be a 99 percent confidence interval based on the same data?

(A) $(0.30, 0.38)$

(B) $(0.35, 0.43)$

(C) $(0.36, 0.48)$

(D) $(0.37, 0.41)$

(E) $(0.40, 0.41)$

29. Luisa is a manager at a pet insurance company with many customers. Sixty percent of the company's customers filed claims in the past year. Luisa will randomly select 12 customers from all the company's customers.

Assuming the customers selected are independent of each other, which of the following is closest to the probability that more than 2 of the 12 customers sampled did <u>not</u> file claims in the past year?

(A) 0.003

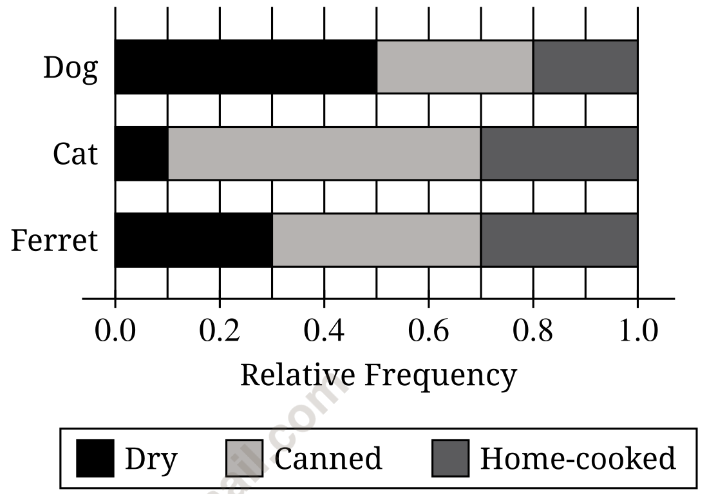
(B) 0.400

(C) 0.407

(D) 0.997

(E) 0.999

30. A veterinarian interested in pet nutrition surveyed owners of a single pet (either a dog, a cat, or a ferret) to ask whether they feed their pet primarily dry pet food, canned pet food, or home-cooked food. The following segmented bar graph shows the distribution of the three types of food for each type of pet.



Which of the following statements about pet owners in the survey is supported by the segmented bar graph?

(A) Of all owners who fed their pet primarily dry pet food, the least number owned a cat.

(B) More than half of the owners fed their pet primarily canned pet food.

(C) Ninety percent of owners fed their pets primarily dry pet food.

(D) The number of ferret owners who fed their pets primarily dry pet food was equal to the number of ferret owners who fed their pets primarily home-cooked food.

(E) A majority of dog owners fed their pets primarily canned pet food.

31. For a middle school science project, Jalen measured the pH of 25 vinegar products. The summary statistics for the values of the pH are shown in the following computer output.

Variabke	N	Mean	Median	StDev	Minimum	Maximum	Q1	Q3
pH	25	2.50	2.49	0.26	2.11	2.93	2.23	2.78

Based on the $1.5 \times$IQR rule for outliers, which of the following statements is true?

(A) The minimum pH value of 2.11 is an outlier, and no other pH value could be identified as an outlier.

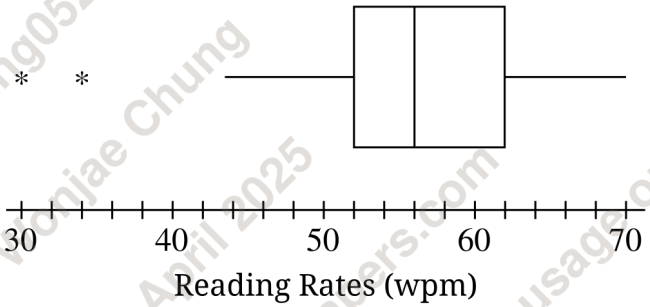
(B) The maximum pH value of 2.93 is an outlier, and no other pH value could be identified as an outlier.

(C) The minimum pH value of 2.11 is an outlier, and there could be other pH values identified as outliers that are below the first quartile, but no pH value above the third quartile could be identified as an outlier.

(D) The maximum pH value of 2.93 is an outlier, and there could be other pH values identified as outliers that are above the third quartile, but no pH value below the first quartile could be identified as an outlier.

(E) There are no pH values that could be identified as outliers.

32. The boxplot summarizes the reading rates, in words per minute (wpm), for 160 elementary school students.



Based on the boxplot, which of the following statements is not true for the data set?

(A) The data set contains two outliers.

(B) The range is greater than 30 wpm.

(C) At least half of the students had reading rates of at most 56 wpm.

(D) More students had reading rates between 30 wpm and 52 wpm than between 52 wpm and 70 wpm.

(E) More students had reading rates greater than 52 wpm than had reading rates less than 56 wpm.

33. A biomedical scientist from a major research university suggested that students should always ask for help when it is needed. A science teacher decided to investigate whether some students are more likely to ask for help than others. The teacher randomly selected 200 students from a Biology II course and gave them directions for a short lab procedure that were purposefully unclear. For each student, the teacher recorded whether the student asked for clarification and the grade the student earned, A, B, or C, in the preceding Biology I course. Results are listed in the following table.

	Asked for Clarification	**Did Not Ask for Clarification**	**Total**
Earned an A in Biology I	20	6	26
Earned an B in Biology I	71	31	102
Earned an C in Biology I	37	35	72
Total	128	72	200

The value of the test statistic for the appropriate chi-square test is approximately 8.25. Assume the conditions for inference were met. Do the sample data provide convincing statistical evidence of an association between grade in the preceding Biology I course and whether a student asks for clarification at the 5 percent level of significance?

(A) Yes, because the p-value is greater than 0.05.

(B) Yes, because the p-value is less than 0.05.

(C) No, because the p-value is greater than 0.05.

(D) No, because the p-value is less than 0.05.

(E) No, because the test statistic is greater than 0.05.

34. The tail length of Siberian tigers is approximately normally distributed with a mean of 0.85 meter and a standard deviation of 0.13 meter.

Which of the following is the best interpretation of the z-score for a particular Siberian tiger with a tail length of 0.8 meter?

(A) The tiger's tail length is approximately 0.38 standard deviation below the mean.

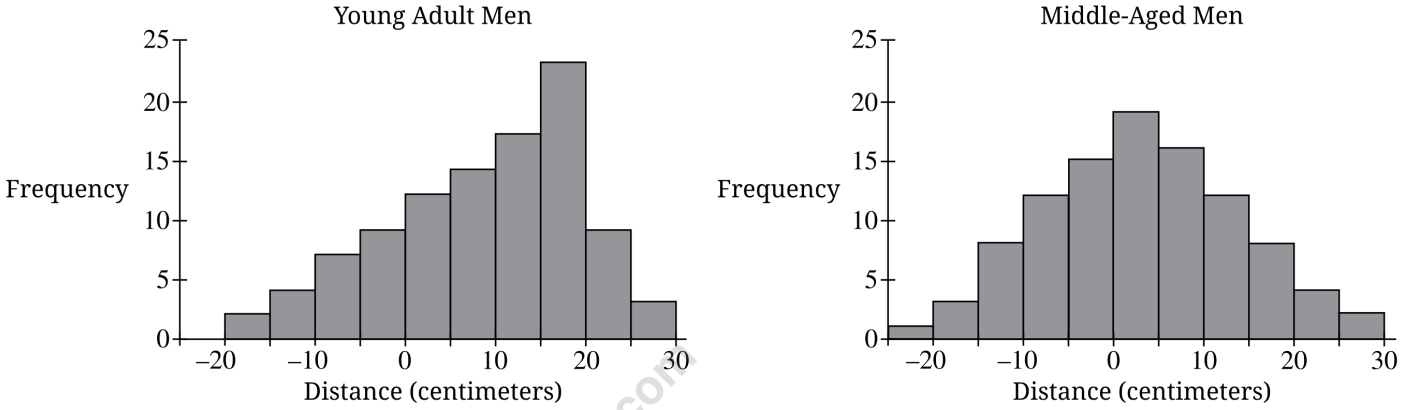
(B) The tiger's tail length is approximately 0.38 standard deviation above the mean.

(C) The tiger has an approximate 0.38 probability of having a tail length of 0.8 meter.

(D) The tiger's tail length is approximately 0.0065 meter greater than the standard deviation.

(E) The tiger's tail length is approximately 0.05 meter below the mean.

35. A common measure of flexibility is the sit-and-reach test, where a person sits on the floor with legs straight in front. The person then reaches as far forward as possible, and the distance reached past the feet is recorded. The histograms summarize the results of the sit-and-reach test for two groups, young adult men and middle-aged men. Negative distances represent a reach that is not as far as the feet.



Which of the following statements is the best description of the distributions of distances?

(A) The distribution for young adult men is approximately symmetric, and the distribution for middle-aged men is skewed to the right.

(B) The distribution for young adult men is approximately symmetric, and the distribution for middle-aged men is skewed to the left.

(C) The distribution for young adult men is skewed to the right, and the distribution for middle-aged men is approximately symmetric.

(D) The distribution for young adult men is skewed to the left, and the distribution for middle-aged men is approximately symmetric.

(E) The distribution for young adult men is skewed to the left, and the distribution for middle-aged men is skewed to the right.

36. The yields, in bushels per acre, of soybeans for 25 randomly selected counties in the midwestern region of the country in 2016 are summarized in the table.

n	$\bar{x}$	s
25	45.88	7.94

Assuming all conditions for inference are met, which of the following is a 99 percent confidence interval for the population mean yield of soybeans for all counties in the midwestern region in 2016?

(A) $45.88 + 2.797\left(\dfrac{7.94}{25}\right)$

(B) $45.88 + 2.797\left(\dfrac{7.94}{5}\right)$

(C) $45.88 + 2.576\left(\dfrac{7.94}{5}\right)$

(D) $45.88 + 2.787\sqrt{\dfrac{7.94}{25}}$

(E) $45.88 + 2.797\sqrt{\dfrac{7.94}{25}}$

37. Ming is learning the game of tennis and will practice serving a tennis ball within bounds. Assume that the probability of Ming serving a tennis ball within bounds is 0.40 and that her serves are independent of each other.

What is the probability that the first time Ming serves a tennis ball within bounds will occur on her 4th attempt?

(A) 0.4

(B) $1-(0.4)^4$

(C) $(0.6)^3(0.4)$

(D) $\binom{4}{1}(0.4)(0.6)^3$

(E) $\binom{4}{1}(0.4)(0.6)^3+\binom{4}{2}(0.4)^2(0.6)^2+\binom{4}{3}(0.4)^3(0.6)$

38. Caleb designed an experiment to investigate whether listening to music or chewing gum affects the time to complete a level of a certain video game. Before the game started, each person in a sample was randomly assigned to either listen to music or not listen to music. Each person was also randomly assigned to either chew gum or not chew gum. At the end of the game, the time to complete the level for each person in the sample was recorded.

Which of the following correctly identifies the response variable of the experiment and whether the variable is categorical or quantitative?

(A) The response variable is the time to complete the level and the variable is quantitative.

(B) The response variable is the time to complete the level and the variable is categorical.

(C) The response variable is whether the person was assigned to listen to music or chew gum and the variable is categorical.

(D) The response variable is whether listening to music or chewing gum affects time to complete the level and the variable is quantitative.

(E) The response variable is whether listening to music or chewing gum affects time to complete the level and the variable is categorical.

39. For a particular video game, Ariana's mean number of points earned per game is 190, and the standard deviation is 40. Suppose that she plays 50 games in one evening, and her goal is to obtain a mean score of at least 200 for the 50 games.

How many standard deviations above 190 is her goal?

(A) 0.25

(B) 1.41

(C) 1.77

(D) 11.18

(E) 12.50

40. The distribution of weight of a small watermelon is known to have a mean of 5 pounds and a standard deviation of 1.2 pounds. Six small watermelons fit in a crate, and the crate weighs exactly 1 pound.

Assuming the weights of small watermelons are independent, what is the standard deviation of the total weight of a random sample of 6 small watermelons and the crate?

(A) $\sqrt{6^2(1.2)^2}$
(B) $1+\sqrt{6^2(1.2)^2}$
(C) $1+6^2(1.2)^2$
(D) $\sqrt{6(1.2)^2}$
(E) $1+\sqrt{6(1.2)^2}$

END OF SECTION I

2 Free-Response Questions

The exam begins on the next page.

STATISTICS

SECTION II

Time—1 hour and 30 minutes

6 Questions

Part A

Suggested Time — 1 hour and 5 minutes

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. A random sample of 25 cans of soda was selected from a factory that produces soda. The amount of soda, in milliliters (ml), contained in each sampled can was measured. The data are shown in the following stem-and-leaf plot.

34	7
35	1 2 3 3 4 4
35	5 5 5 6 7 9
36	2
36	5 6 6 7 8 8
37	0 3 3 4
37	5

Key: 34 | 7 represents 347 ml of soda

(a) Describe the distribution of amount of soda in a can.

The factory uses two machines to fill the soda cans: Machine A and Machine B. The following back-to-back stem-and-leaf plot shows the soda amounts from the same random sample categorized by machine.

Machine A							Machine B					
						34	7					
					4	35	1	2	3	3	4	
					5	35	5	5	6	7	9	
					2	36						
8	8	7	6	6	5	36						
		4	3	3	0	37						
					5	37						

Key: 34 | 7 represents 347 ml of soda

(b) Calculate the median and IQR for each machine.

- Machine A median:
- Machine A IQR:
- Machine B median:
- Machine B IQR:

(c) Compare the variability of the measurements in the amount of soda filled per can by the two machines by referencing the appropriate measure calculated in part (b). Justify your answer.

2. An educational researcher has noticed that because many students can type faster than they can write, lecture notes taken by students using laptops tend to look like a transcript of the instructor's lecture, while handwritten lecture notes tend to be shorter summaries of the key points made during the lecture. The researcher wants to investigate whether the difference in note-taking methods has an impact on student quiz performance and is considering the following two proposals for the investigation.

   - Proposal I: The researcher will select a random sample of students from a list of volunteers and allow the students to use their preferred note-taking method, laptops or handwritten.
   - Proposal II: The researcher will randomly assign the note-taking method (laptops or handwritten) to each student in a group of volunteers, so that half of the volunteers are assigned to each method.

   (a) Which of the two proposals is not an experiment? Justify your answer.

   (b) Could either of the two proposals be used to establish a cause-and-effect relationship? Justify your answer.

   (c) Describe how student study habits could be a confounding variable when using Proposal I for the investigation.

3. In the western United States, water is sometimes diverted from mountain streams and rivers to be used for agriculture and municipal water supplies. Diverting too much water raises the water temperature in the mountain streams and rivers, affecting the fish that live there.

   A certain mountain river is abundant in brown trout and has water diverted from it for city use. Assume that the water temperature along all locations of the river has a distribution with mean 18°C and standard deviation of 2°C.

   A sport-fishing group plans to measure the water temperature at 64 randomly selected locations along the river. Let the random variable F represent the mean water temperature, in degrees Celsius, of 64 locations along the river selected at random by the sport-fishing group.

   (a) Describe the distribution of F.

   (b) Fish, such as brown trout, become stressed once water temperatures exceed 18.3°C. Determine the probability F will exceed 18.3°C. Show your work.

   A large metropolitan water district that uses water diverted from the same river is also planning to measure the water temperature at 100 randomly selected locations along the river. Because of sampling variability, the sample collected by the water district is likely to yield a sample mean water temperature different from the sample mean water temperature collected by the sport-fishing group. Let the random variable D represent the mean water temperature, in degrees Celsius, of 100 locations along the river selected at random by the large metropolitan water district.

   (c) Describe the distribution of $D - F$.

4. A veterinarian is comparing two diets, Diet J and Diet W, designed to help overweight large-breed dogs lose weight. Thirty-two overweight large-breed dogs were randomly assigned to one of the two diets, with 16 dogs in each group. The dogs remained on the assigned diet for six months and the weight loss, in pounds, was recorded for each dog. Summary statistics of the weight loss for the two groups are shown.

	Mean (pounds)	**Standard Deviation (pounds)**
Diet J	7.523	0.819
Diet W	6.379	1.012

Assume the distribution of weight loss for all overweight large-breed dogs who use Diet J is approximately normal and assume the distribution of weight loss for all overweight large-breed dogs who use Diet W is approximately normal. Do the data provide convincing statistical evidence, at the significance level of $\alpha = 0.05$, that Diet J is more effective than Diet W in terms of mean weight loss for overweight large-breed dogs that are similar to those used in the study? Complete the appropriate inference procedure to support your answer.

5. A survey conducted by a national research center in 2018 asked a random sample of people over age 12 years old in the United States whether it is very important to them personally to have a lot of money. Responses are summarized based on age in the following table of counts.

	Adults	**Teenagers**
Very important	281	86
Not very important	180	168
Total	461	254

(a) Construct a graph to compare the responses of adults and teenagers.

(b) A hypothesis test will be conducted to investigate whether the proportion of all United States adults who would respond that it is very important to have a lot of money is different from the proportion of all United States teenagers who would respond that it is very important to have a lot of money. Identify the appropriate test and state the hypotheses.

(c) The p-value of the hypothesis test was found to be less than 0.0001. Based on the p-value, what is the appropriate conclusion for the test? Justify your answer.

STATISTICS

SECTION II, Part B

Suggested Time — 25 minutes

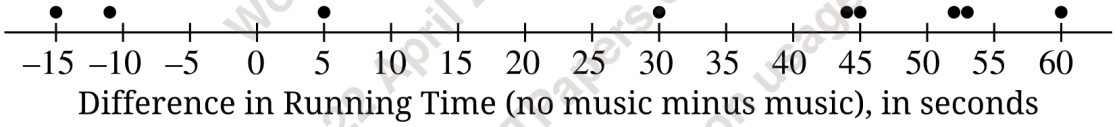
1 Question

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

6. Nine experienced runners participated in a well-designed controlled experiment to investigate whether listening to music while running will allow a runner to run faster. On each of two consecutive Saturdays, the runners ran on a track until they completed 3.1 miles. Researchers recorded the time, in seconds, for each runner to complete

   3.1 miles. On one day the runners listened to music while running, and on the other day the runners did not listen to music. For each runner, a coin was flipped to determine on which day, the first or second, music was used.

   The following dotplot shows the differences in running times (no music minus music) for the 9 runners.



   (a) For the investigation, what does the positive difference of 60 indicate?

If listening to music has no effect on running time, then running time is equally likely to be greater for either condition, music or no music. Under that assumption, the probability that the difference (no music minus music) will be positive for any particular runner is $p = 0.5$. Let X represent the number of runners out of 9 runners who have a positive difference (no music minus music) in running time.

(b) Assume that the differences in running times are independent of one another. From the sample of 9 runners, 7 have a positive difference (no music minus music).

i. If listening to music has no effect on running time, what is the probability that among 9 runners, 7 or more runners will have a positive difference in running times? Show your work.

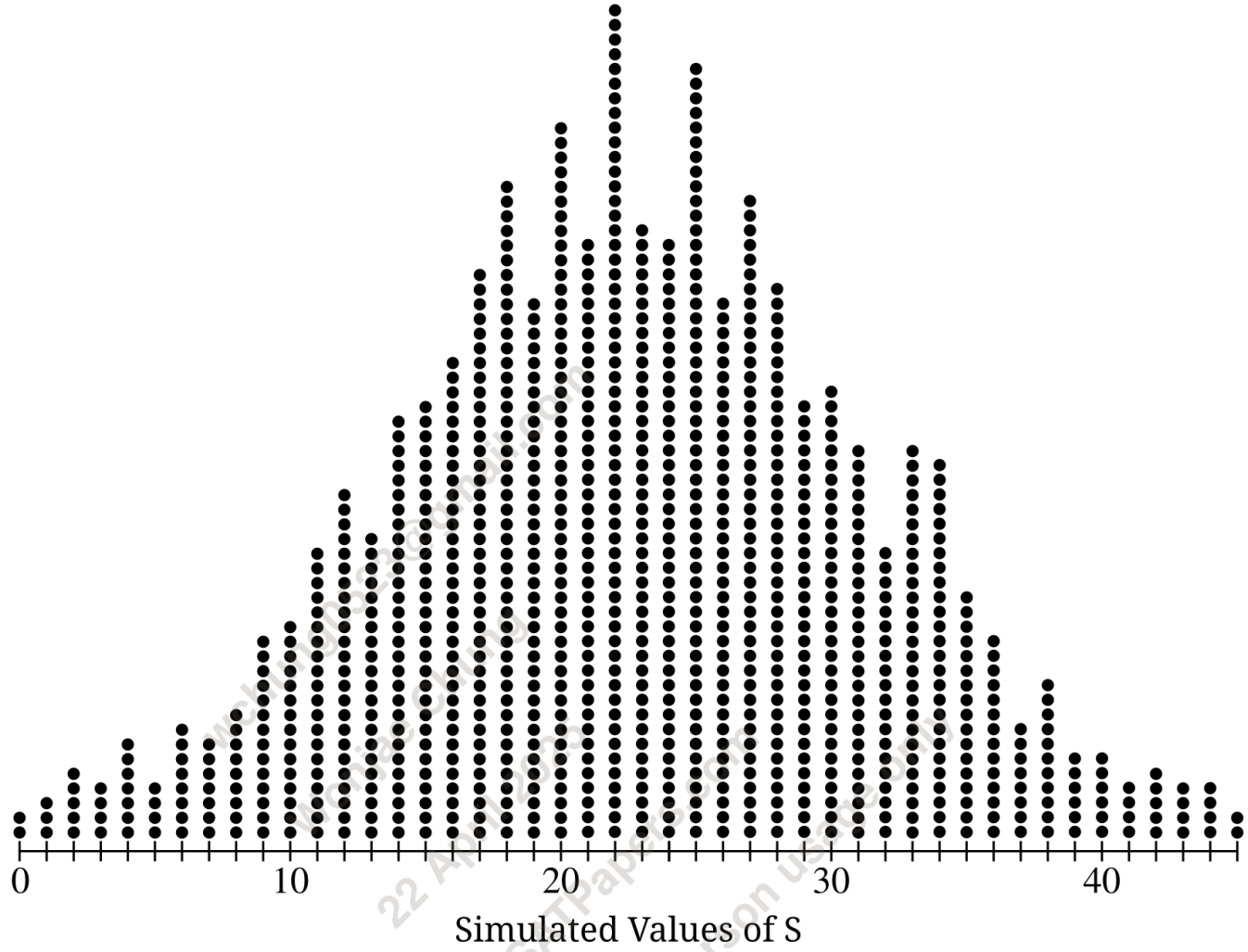
ii. Based on your answer in part (b-i), is it reasonable to believe that runners tend to have longer running times while not listening to music? Explain your answer without performing an inference procedure.

In part (b-ii), only the signs of the differences (positive or negative), not the magnitudes, are used. An alternative method that takes into account both the sign and the magnitude of the differences in running times is based on the sum, S, of ranks corresponding to positive differences. For the experiment described above, S is computed by first ordering the absolute values of the nine differences in times from smallest to largest. Next, ranks are computed for the absolute values of the differences with rank 1 corresponding to the smallest absolute value, rank 2 corresponding to the second smallest absolute value, up to rank 9 corresponding to the largest absolute value. Finally, the value of S is computed by summing only the values of the ranks corresponding to positive differences; the values of the ranks corresponding to negative differences are not included in the sum. The following table displays the differences and the ranks of the absolute values of the differences for the nine runners who participated in the experiment, where rank 1 indicates the smallest absolute value and rank 9 indicates the largest absolute value.

Differences	5	-11	-15	30	44	45	52	53	60
Ranks of absolute values of the differences	1	2	3	4	5	6	7	8	9

(c) Determine the value of S.

A simulation was carried out to investigate the sampling distribution of S. The value of S was calculated for each of 1,000 trials of the simulation, where each trial was simulated under the assumption that listening to music has no effect on running time. The results are shown in the dotplot below.



(d) Based on the simulation results and your response to part (c), is it reasonable to conclude that not listening to music tends to result in longer running times? Justify your conclusion.

STOP
END OF EXAM

3 Answer for Multiple Choice Questions

Question	Answer	Question	Answer	Question	Answer	Question	Answer
1	D	11	B	21	A	31	E
2	E	12	B	22	D	32	D
3	B	13	C	23	E	33	B
4	C	14	D	24	E	34	A
5	C	15	D	25	C	35	D
6	D	16	B	26	E	36	B
7	A	17	D	27	B	37	C
8	E	18	D	28	B	38	A
9	D	19	E	29	C	39	C
10	B	20	A	30	D	40	D

4 Scoring Guidelines for Free-Response Questions

2022 Form I

1. A university that offers evening classes to a large number of students who commute to campus wants to estimate the mean commute time of all evening students. A parking facilities attendant selects one parking lot on campus, then selects the first 60 students who drive into the lot on a certain evening. Each of the 60 students is asked to provide their typical commute time to campus. The university will create a confidence interval for the mean commute time of evening students.

Are the conditions for inference satisfied?

(A) Yes, all conditions for inference are satisfied.

(B) No, because the distribution of the population mean commute time is not known to be normally distributed.

(C) No, because the sample size is greater than 10% of the population size.

(D) No, because the students selected are not a random sample of all evening students. ✓

(E) No, because the sample size of 60 is not large enough to assume that the sampling distribution of the sample mean commute time is approximately normal.

2. The distribution of the length of employment for workers at a factory is moderately skewed to the right, with mean 7.4 years and standard deviation 5.6 years. The factory manager will select a random sample of 100 employees to survey.

Which of the following best describes the sampling distribution of the sample means of length of employment for samples of size 100?

(A) Moderately skewed to the right, with mean 7.4 years and standard deviation 5.6 years

(B) Moderately skewed to the right, with mean 7.4 years and standard deviation less than 5.6 years

(C) Moderately skewed to the right, with mean less than 7.4 years and standard deviation less than 5.6 years

(D) Approximately normal, with mean less than 7.4 years and standard deviation 5.6 years

(E) Approximately normal, with mean 7.4 years and standard deviation less than 5.6 years ✓

3. Researchers from a national research organization investigated which platform (television, social media, newspaper, Internet, etc.) United States adults prefer to use to obtain their news. In 2016, of the 200 randomly selected United States adults that were surveyed, 114 selected television as their preferred platform. In 2018, of the 200 randomly selected United States adults that were surveyed, 98 selected television as their preferred platform. The conditions for inference are met. A 95 percent confidence interval for the difference in the population proportions of United States adults that use television as their preferred platform for obtaining news between 2016 and 2018 (2016 minus 2018) was found to be $(-0.02, 0.18)$.

Based on the interval, is it reasonable for the researchers to claim that there is convincing statistical evidence that the population proportion of United States adults that use television as their preferred method to obtain news differed between 2016 and 2018 ?

(A) No, because the researchers should have used a greater confidence level to obtain convincing evidence.

(B) No, because zero is included in the interval there is not convincing statistical evidence that the population proportion differed between 2016 and 2018. ✓

(C) Yes, because zero is included in the interval there is convincing statistical evidence that the population proportion differed between 2016 and 2018.

(D) Yes, because most of the interval is positive there is convincing statistical evidence that the population proportion increased from 2016 to 2018.

(E) Yes, because 0.57 is greater than 0.49 there is convincing statistical evidence that the population proportion decreased from 2016 to 2018.

4. A solar power engineer took a random sample of houses and installed the same type of solar panels using two different methods on each house to investigate whether there is a mean difference in the angles of installation between the two methods for all houses in the population of interest. The engineer found the sample mean difference between the two methods to be 0.2 degree and the p-value for a two-sided matched-pairs t-test for the mean difference to be 0.65.

Assuming the conditions for inference are met, which of the following statements is the best interpretation of the p-value?

(A) The probability that the null hypothesis is true is 0.65.

(B) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of 0.2 degree or -0.2 degree.

(C) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of greater than 0.2 degree or less than -0.2 degree. ✓

(D) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of greater than 0.2 degree.

(E) If the null hypothesis is true, the probability is 0.65 of observing a mean difference of less than -0.2 degree.

5. Each accountant at a large accounting firm was classified according to accountant level (junior or senior) and method of transportation to work (walk, bus, drive alone, or carpool). The responses of the 320 accountants at the firm are summarized in the table.

	Junior	**Senior**	**Total**
Walk	25	3	28
Bus	87	12	99
Drive alone	96	25	121
Carpool	52	20	72
Total	260	60	320

What proportion of the accountants at the firm are at the senior level and carpool to work?

2022 Form I

(A) $\frac{20}{60}$

(B) $\frac{20}{72}$

(C) $\frac{20}{320}$ ✓

(D) $\frac{112}{320}$

(E) $\frac{132}{320}$

6. The following table shows the classification of all 51 dogs from an animal shelter by whether the dogs lived mostly outdoors or mostly indoors before coming to the shelter and whether they tested positive or negative for a certain skin condition.

	Positive	**Negative**	**Total**
Mostly Outdoors	10	20	30
Mostly Indoors	8	13	21
Total	18	33	51

One dog from the animal shelter will be selected at random. Based on the table, which of the following statements is correct?

(A) The probability of selecting a dog that tested positive is greater than 0.5.

(B) The probability of selecting a dog that lived mostly indoors is greater than 0.5.

(C) The probability of selecting a dog that tested positive is greater than the probability of selecting a dog that tested negative.

(D) The probability of selecting a dog that tested positive given that the dog lived mostly outdoors is less than the probability of selecting a dog that tested positive given that the dog lived mostly indoors. ✓

(E) The probability of selecting a dog that lived mostly outdoors given that the dog tested positive is greater than the probability of selecting a dog that lived mostly outdoors given that the dog tested negative.

7. Researchers are examining the relationship between hours of sleep and athletic performance among college athletes. Athletic performance will be measured on a numeric scale, with greater numbers indicating better performance. The researchers expect that the more hours the athletes sleep, the better they will perform. Assuming all conditions for inference are met, the researchers will create a 95 percent confidence interval for the slope of the regression line for predicting athletic performance from amount of sleep.

For which of the following would the confidence interval support the researchers' expectations?

(A) The confidence interval includes only positive values. ✓

(B) The confidence interval includes only negative values.

(C) The confidence interval has a width less than 1.

(D) The confidence interval has a width greater than 1.

(E) The confidence interval includes the value 0.

2022 Form I

8. A company that sells baby food is interested in the baby food preferences of all families with toddlers from a certain city. A representative from the company sets up a booth at one grocery store in the city that will be used to investigate baby food preference. The representative will ask people with toddlers who walk past the booth to complete the survey.

Which of the following best describes the sampling method the company will use?

(A) Systematic random sampling

(B) Simple random sampling

(C) Cluster sampling

(D) Stratified random sampling

(E) Convenience sampling ✓

9. A large online retailer places packed and ready-to-be-shipped boxes in a line on a conveyor belt. A sample of boxes will be used to estimate the mean weight of boxes shipped on a given day. One proposed method for selecting the sample is as follows:

- Generate a random integer from 1 to 20.
- Use that integer to select an initial box from the first 20 boxes in line on the conveyor belt.
- Select every 25th box that follows the initial box on the conveyor belt for the rest of the day.

Which of the following best describes the proposed sampling method?

(A) Stratified random sample, stratified by time of day

(B) Cluster sample, with times of day as clusters

(C) Cluster sample, with box weights as clusters

(D) Systematic random sample ✓

(E) Convenience sample

2022 Form I

10. Each voter from a random sample of 334 registered voters was asked their impression of two candidates running for the same national office. The table summarizes the responses.

	Candidate A	**Candidate B**
Favorable	138	200
Unfavorable	44	47
No opinion	88	47
Have not heard of	64	40

Which of the following is the most appropriate method to use to estimate the proportion of all registered voters who have a favorable impression of Candidate A?

(A) A one-sample z-interval for estimating a sample proportion

(B) A one-sample z-interval for estimating a population proportion

(C) A matched-pairs t-interval for estimating a mean difference

(D) A two-sample z-interval for estimating a difference between sample proportions

(E) A two-sample z-interval for estimating a difference between population proportions

11. A raffle for a charity fund-raiser is being planned. Each of 2000 raffle tickets will be sold for $\$1.00$. The holders of 32 winning tickets will each win a prize. The table shows the prize values and the number of prizes for each value.

Prize Value	**Number of Prizes**
$25	20
$50	10
$300	2

The random variable W represents the value of the prize won for a single ticket minus the cost of the ticket. What is the expected value of W?

(A) $-\$1.00$

(B) $-\$0.20$

(C) $\$0.80$

(D) $\$49.00$

(E) $\$124.00$

2022 Form I

12. In a study researching how donating to charity can affect a person's happiness, 96 participants were given $\$5$ a day for one week. Each participant was randomly assigned to one of two groups. Those assigned to the first group were asked to spend the money on themselves, and those assigned to the second group were asked to donate the money to charity. At the end of the week, all of the participants were asked to rate their overall level of happiness on a scale from 0 to 100, with higher scores indicating greater levels of happiness.

Which of the following are the appropriate null and alternative hypotheses to test whether the sample data provide convincing statistical evidence that donating to charity results in greater happiness than spending on oneself, on average?

(A) $H_0: \mu_{self} - \mu_{charity} = 0$
$H_a: \mu_{self} - \mu_{charity} > 0$

(B) $H_0: \mu_{self} - \mu_{charity} = 0$
$H_a: \mu_{self} - \mu_{charity} < 0$ ✓

(C) $H_0: \mu_{self} - \mu_{charity} = 0$
$H_a: \mu_{self} - \mu_{charity} \neq 0$

(D) $H_0: \mu_{self} - \mu_{charity} < 0$
$H_a: \mu_{self} - \mu_{charity} = 0$

(E) $H_0: \mu_{self} - \mu_{charity} > 0$
$H_a: \mu_{self} - \mu_{charity} < 0$

13. Suppose a hypothesis test will be used to investigate whether the proportion of United States adults who have been a victim of a cybercrime, such as identity theft, is greater than 0.25. Suppose that the hypothesis test is conducted using a significance level of 0.05 and the power of the test is determined for a specific alternative value using the significance level of 0.05.

If the significance level of the hypothesis test is changed to 0.01 and the power of the test is computed for the same alternative value using a significance level of 0.01, which of the following best describes the change(s) to the probability of a Type I error and the power of the test?

(A) The probability of a Type I error would decrease, and the power of the test would stay the same.

(B) The probability of a Type I error would decrease, and the power of the test would increase.

(C) The probability of a Type I error would decrease, and the power of the test would decrease. ✓

(D) The probability of a Type I error would increase, and the power of the test would increase.

(E) The probability of a Type I error would increase, and the power of the test would decrease.

2022 Form I

14. A large company has operated assuming that the rate of absenteeism is the same for each day of the week. The new human resources director at the company claims that the assumption is false and that the rate of absenteeism is different for some days of the week as compared with others. To test the claim, a random sample of 150 absences was selected, and the day of the week on which the absence occurred was recorded. The counts are shown in the following table.

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	40	27	21	34	28

The appropriate chi-square test will be performed to test the claim.

What is the contribution of the Monday absences to the calculation of the chi-square test statistic?

(A) 0.333

(B) 2.500

(C) 3.000

(D) 3.333 ✓

(E) 10.000

15. The manager of a restaurant plans to survey a random sample of customers to estimate the proportion of customers who would return within a week if given a coupon for a free dessert to use on their next visit. A similar coupon was given last year, and 30 percent of the customers who were given the coupon returned within a week. The manager will use the observed proportion from last year as a preliminary estimate of the population proportion and wants the margin of error to be at most 0.04 at a 90 percent level of confidence.

Of the following, which is the least number of customers who should be surveyed such that the estimate of the margin of error of the 90 percent confidence interval will be no greater than 0.04?

(A) 15

(B) 260

(C) 310

(D) 370 ✓

(E) 610

2022 Form I

16. An experiment was conducted to investigate whether submersion in cold water causes a lower heart rate. The experiment used 50 volunteers. The 25 youngest volunteers had their heart rate measured while holding their breath for 30 seconds with their face submerged in cold water. The 25 oldest volunteers had their heart rate measured while holding their breath for 30 seconds with their face not submerged in cold water. The mean heart rate for volunteers who had their face submerged in cold water was lower than the mean heart rate for volunteers who did not have their face submerged in cold water.

Which of the following elements of a well-designed experiment is missing?

I. Comparison of at least two treatment groups
II. Random assignment of treatments to experimental units
III. Replication

(A) I only

(B) II only ✓

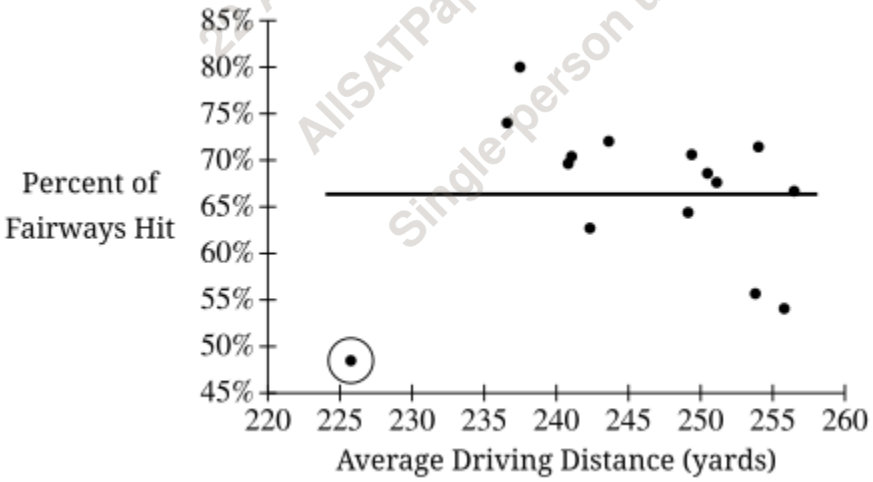
(C) I and II only

(D) II and III only

(E) I, II, and III

17. A sample of 15 golfers who played a golf course on a certain day was selected. For each golfer, the average driving distance (x), in yards, and the percent of fairways hit on the drive (y) were recorded.

The scatterplot displays the percent of fairways hit versus the average driving distance. Also shown is the least-squares regression line, $\hat{y} = 66.228 + 0.0002x$.



The point circled on the scatterplot is considered an influential point. A new least-squares regression line will be calculated with the influential point removed. How will the removal of the influential point affect the new least-squares regression line for the remaining 14 points?

2022 Form I

(A) The y-intercept will remain the same, and the slope will be negative.

(B) The y-intercept will decrease, and the slope will be negative.

(C) The y-intercept will decrease, and the slope will be positive.

(D) The y-intercept will increase, and the slope will be negative. ✓

(E) The y-intercept will increase, and the slope will be positive.

18. For the population of people in the United States who are at least 18 years of age, approximately 61 percent voted in a recent election. A random sample of 200 people and a random sample of 100 people will be selected from the population.

For which sample size is it more likely that greater than 70 percent of the sample voted in the recent election?

(A) They are equally likely, because the mean of the sampling distribution is the same for both sample sizes.

(B) It is more likely for a sample size of 200 people, because there is more variability in the sampling distribution for larger sample sizes.

(C) It is more likely for a sample size of 200 people, because there is less variability in the sampling distribution for larger sample sizes.

(D) It is more likely for a sample of size 100 people, because there is more variability in the sampling distribution for smaller sample sizes. ✓

(E) It is more likely for a sample size of 100 people, because there is less variability in the sampling distribution for smaller sample sizes.

19. The probability that a particular electrical component operates successfully is $\frac{3}{4}$. An electrical system consists of three such components that operate independently. The system will function successfully if at least one of the three components operates successfully.

What is the probability that the system will function successfully?

(A) $\frac{1}{64}$

(B) $\frac{27}{64}$

(C) $\frac{37}{64}$

(D) $\frac{3}{4}$

(E) $\frac{63}{64}$ ✓

20. A food scientist conducting a study to investigate the nutritional habits of teenagers in a city selected a random sample of 100 teenagers in the city. Each of the 100 teenagers was asked to complete a survey with questions about age and the number of ounces of carbonated drinks consumed on the previous day.

Assuming the conditions for inference have been met, which of the following is an appropriate test and alternative hypothesis to use to investigate whether the average number of ounces of carbonated drinks consumed in a day increases with age for teenagers in the city?

2022 Form I

- (A) A t-test for the slope of a regression model with alternative hypothesis: $\mathrm{H_a}: \ \beta > 0$ ✓
- (B) A t-test for the slope of a regression model with alternative hypothesis: $\mathrm{H_a}: \ \beta < 0$
- (C) A t-test for the slope of a regression model with alternative hypothesis: $\mathrm{H_a}: \ \beta = 0$
- (D) A one-sample z-test for a population proportion with alternative hypothesis: $\mathrm{H_a}: \ p > 0.5$
- (E) A two-sample t-test for a difference of two population means with alternative hypothesis: $\mathrm{H_a}: \ \mu_{Older} \neq \mu_{Younger}$

21. For a random sample of 50 newborn Holstein calves from the United States, the mean weight was 80 pounds and the standard deviation was 6 pounds. In order to compare these findings with those of other countries, an agricultural engineer converted the measured calf weights from pounds to kilograms by multiplying each weight by 0.454.

What are the mean and standard deviation of the converted weights of the 50 newborn Holstein calves?

- (A) The mean weight is 36.32 kilograms and the standard deviation is 2.724 kilograms. ✓
- (B) The mean weight is 36.32 kilograms and the standard deviation is 6 kilograms.
- (C) The mean weight is 80 kilograms and the standard deviation is 2.724 kilograms.
- (D) The mean weight is 80 kilograms and the standard deviation is 6 kilograms.
- (E) The mean weight is approximately 176.21 kilograms and the standard deviation is 6 kilograms.

22. The distribution of housing prices in a city includes several outliers at the upper end of the distribution and no outliers at the lower end of the distribution.

Which of the following is the most resistant, or robust, measure of center of the distribution of housing prices?

- (A) The mean, because it is affected by the outliers.
- (B) The mean, because it is not greatly affected by the outliers.
- (C) The median, because it is affected by the outliers.
- (D) The median, because it is not greatly affected by the outliers. ✓
- (E) The mode, because the data are categorical.

23. A study reported that 14 percent of cars registered in 2015 had a manual transmission. An automotive researcher collected data on cars registered in 2021 to investigate whether the proportion of registered cars with a manual transmission is less than 0.14. The researcher took a random sample of 150 cars registered in 2021 and found 18 of the cars had a manual transmission.

After ensuring the conditions for inference were met, the researcher conducted an appropriate hypothesis test and calculated a p-value of 0.24. At the 5 percent level of significance, which of the following conclusions should be made?

2022 Form I

(A) The null hypothesis should be rejected. There is significant evidence that the population proportion of registered cars with a manual transmission in 2021 is less than 0.14.

(B) The null hypothesis should be rejected. There is not significant evidence that the population proportion of registered cars with a manual transmission in 2021 is less than 0.14.

(C) The null hypothesis should not be rejected. There is significant evidence that the population proportion of registered cars with a manual transmission in 2021 is not less than 0.14.

(D) The null hypothesis should not be rejected. There is significant evidence that the population proportion of registered cars with a manual transmission in 2021 is equal to 0.14.

(E) The null hypothesis should not be rejected. There is not significant evidence that the population proportion of registered cars with a manual transmission in 2021 is less than 0.14. ✓

24. For which of the following distributions is the proportion of values between 900 and 1,000 the greatest?

(A) A normal distribution with mean 900 and standard deviation 100

(B) A normal distribution with mean 950 and standard deviation 100

(C) A normal distribution with mean 800 and standard deviation 100

(D) A normal distribution with mean 1,000 and standard deviation 400

(E) A normal distribution with mean 950 and standard deviation 50 ✓

25. An owner of a novelty business operates both an online store and a retail store. The owner took a random sample of 50 sales from each store during the past year and recorded the purchase amount of each sale. Summary statistics are shown in the following table.

	Online	**Retail**
n	50	50
Mean	$42.80	$27.50
Standard deviation	$10.50	$8.25

Assuming the conditions for inference were met, the owner calculated a 95 percent confidence interval for the difference in mean purchase amounts (online minus retail) as ($11.55, $19.05).

Which of the following is an appropriate interpretation of the 95 percent confidence interval?

2022 Form I

(A) We are 95% confident that the purchase amount from a randomly selected online store sale will be between \$11.55 and \$19.05 greater than the purchase amount from a randomly selected retail store sale.

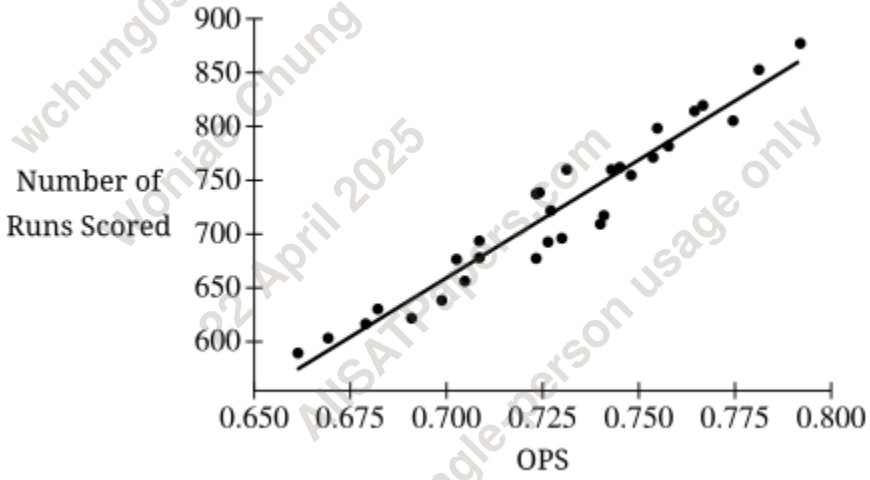
(B) We are 95% confident that the sample mean purchase amount from the online store is between \$11.55 and \$19.05 greater than the sample mean purchase amount from the retail store for the past year.

(C) We are 95% confident that the population mean purchase amount from the online store is between \$11.55 and \$19.05 greater than the population mean purchase amount from the retail store for the past year. ✓

(D) We are 95% confident that the population mean purchase amount for the past year from both stores combined is between \$11.55 and \$19.05.

(E) We are 95% confident that if new samples of the same size were taken from each store, the sample mean purchase amount from the online store would be between \$11.55 and \$19.05 greater than the sample mean purchase amount from the retail store.

26. On-base percentage plus slugging (OPS) is a statistic used in baseball to measure a team's batting success. The number of runs scored and OPS for 30 baseball teams was used to conduct a linear regression analysis. The scatterplot and computer output for the regression analysis is shown.



Term	Coef	SE Coef	T-Value	P-Value
Constant	−838.40	77.99	−10.75	< 0.0001
OPS	2144.3	107.1	20.01	< 0.0001
$S = 19.516$		R-Sq $= 93.47\%$		R-Sq(adj) $= 93.23\%$

Which of the following is the most appropriate interpretation of the statistic 93.47% in the regression output?

2022 Form I

(A) There is a strong, positive, linear relationship between number of runs scored and OPS.

(B) The typical deviation between observed and predicted number of runs scored is 0.9347.

(C) For each one-unit increase in OPS, the regression model predicts an increase of 93.47 runs scored.

(D) 93.47% of the observed number of runs scored are close to the regression line.

(E) 93.47% of the variation in number of runs scored can be explained by the linear regression with OPS. ✓

27. A paint manufacturer conducted an experiment to investigate whether a new formula produces paint that lasts longer on the exterior of a building than paint produced using an older formula. A grid of 20 squares was drawn on one side of an exterior wall of a building. The new formula was randomly assigned to 10 squares, and the older formula was assigned to the remaining 10 squares. Each square was painted with the assigned formula. One year later, each square was rated on how well the paint had lasted, using a numerical scale for which larger values indicated longer-lasting paint.

Which of the following best describes the design of the experiment?

(A) A completely randomized design, because the two formulas were compared.

(B) A completely randomized design, because each formula was randomly assigned to 10 squares. ✓

(C) A matched pairs design, because the same number of squares were assigned to each formula.

(D) A matched pairs design, because two formulas were compared.

(E) A randomized block design, because the wall was partitioned into 20 squares.

28. A 95 percent confidence interval for the proportion of parents who use parental controls for blocking, filtering, or monitoring their teenagers' online activities is $(0.36, 0.42)$.

Which of the following could be a 99 percent confidence interval based on the same data?

(A) $(0.30, 0.38)$

(B) $(0.35, 0.43)$ ✓

(C) $(0.36, 0.48)$

(D) $(0.37, 0.41)$

(E) $(0.40, 0.41)$

29. Luisa is a manager at a pet insurance company with many customers. Sixty percent of the company's customers filed claims in the past year. Luisa will randomly select 12 customers from all the company's customers.

Assuming the customers selected are independent of each other, which of the following is closest to the probability that more than 2 of the 12 customers sampled did <u>not</u> file claims in the past year?

2022 Form I

(A) 0.003

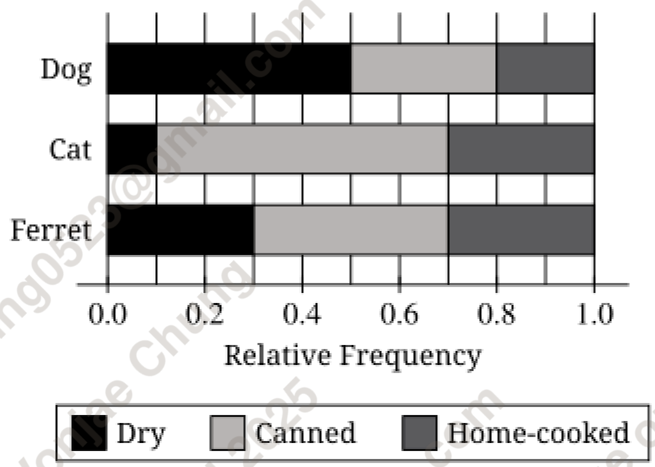
(B) 0.400

(C) 0.917 ✓

(D) 0.997

(E) 0.999

30. A veterinarian interested in pet nutrition surveyed owners of a single pet (either a dog, a cat, or a ferret) to ask whether they feed their pet primarily dry pet food, canned pet food, or home-cooked food. The following segmented bar graph shows the distribution of the three types of food for each type of pet.



Which of the following statements about pet owners in the survey is supported by the segmented bar graph?

(A) Of all owners who fed their pet primarily dry pet food, the least number owned a cat.

(B) More than half of the owners fed their pet primarily canned pet food.

(C) Ninety percent of owners fed their pets primarily dry pet food.

(D) The number of ferret owners who fed their pets primarily dry pet food was equal to the number of ferret owners who fed their pets primarily home-cooked food. ✓

(E) A majority of dog owners fed their pets primarily canned pet food.

31. For a middle school science project, Jalen measured the pH of 25 vinegar products. The summary statistics for the values of the pH are shown in the following computer output.

Variable	N	Mean	Median	StDev	Minimum	Maximum	Q1	Q3
pH	25	2.50	2.49	0.26	2.11	2.93	2.23	2.78

Based on the $1.5 \times \text{IQR}$ rule for outliers, which of the following statements is true?

2022 Form I

(A) The minimum pH value of 2.11 is an outlier, and no other pH value could be identified as an outlier.

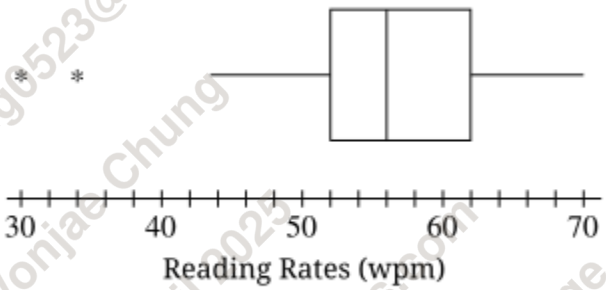
(B) The maximum pH value of 2.93 is an outlier, and no other pH value could be identified as an outlier.

(C) The minimum pH value of 2.11 is an outlier, and there could be other pH values identified as outliers that are below the first quartile, but no pH value above the third quartile could be identified as an outlier.

(D) The maximum pH value of 2.93 is an outlier, and there could be other pH values identified as outliers that are above the third quartile, but no pH value below the first quartile could be identified as an outlier.

(E) There are no pH values that could be identified as outliers. ✓

32. The boxplot summarizes the reading rates, in words per minute (wpm), for 160 elementary school students.



Based on the boxplot, which of the following statements is not true for the data set?

(A) The data set contains two outliers.

(B) The range is greater than 30 wpm.

(C) At least half of the students had reading rates of at most 56 wpm.

(D) More students had reading rates between 30 wpm and 52 wpm than between 52 wpm and 70 wpm. ✓

(E) More students had reading rates greater than 52 wpm than had reading rates less than 56 wpm.

33. A biomedical scientist from a major research university suggested that students should always ask for help when it is needed. A science teacher decided to investigate whether some students are more likely to ask for help than others. The teacher randomly selected 200 students from a Biology II course and gave them directions for a short lab procedure that were purposefully unclear. For each student, the teacher recorded whether the student asked for clarification and the grade the student earned, A, B, or C, in the preceding Biology I course. Results are listed in the following table.

	Asked for Clarification	Did Not Ask for Clarification

2022 Form I

			Total
Earned an A in Biology I	20	6	26
Earned a B in Biology I	71	31	102
Earned a C in Biology I	37	35	72
Total	128	72	200

The value of the test statistic for the appropriate chi-square test is approximately 8.25. Assume the conditions for inference were met. Do the sample data provide convincing statistical evidence of an association between grade in the preceding Biology I course and whether a student asks for clarification at the 5 percent level of significance?

(A) Yes, because the p-value is greater than 0.05.

(B) Yes, because the p-value is less than 0.05. ✓

(C) No, because the p-value is greater than 0.05.

(D) No, because the p-value is less than 0.05.

(E) No, because the test statistic is greater than 0.05.

34. The tail length of Siberian tigers is approximately normally distributed with a mean of 0.85 meter and a standard deviation of 0.13 meter.

Which of the following is the best interpretation of the z-score for a particular Siberian tiger with a tail length of 0.8 meter?

(A) The tiger's tail length is approximately 0.38 standard deviation below the mean. ✓

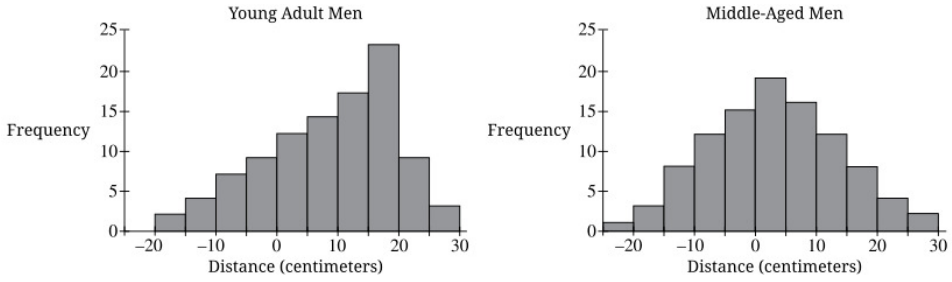
(B) The tiger's tail length is approximately 0.38 standard deviation above the mean.

(C) The tiger has an approximate 0.38 probability of having a tail length of 0.8 meter.

(D) The tiger's tail length is approximately 0.0065 meter greater than the standard deviation.

(E) The tiger's tail length is approximately 0.05 meter below the mean.

35. A common measure of flexibility is the sit-and-reach test, where a person sits on the floor with legs straight in front. The person then reaches as far forward as possible, and the distance reached past the feet is recorded. The histograms summarize the results of the sit-and-reach test for two groups, young adult men and middle-aged men. Negative distances represent a reach that is not as far as the feet.



Which of the following statements is the best description of the distributions of distances?

2022 Form I

(A) The distribution for young adult men is approximately symmetric, and the distribution for middle-aged men is skewed to the right.

(B) The distribution for young adult men is approximately symmetric, and the distribution for middle-aged men is skewed to the left.

(C) The distribution for young adult men is skewed to the right, and the distribution for middle-aged men is approximately symmetric.

(D) The distribution for young adult men is skewed to the left, and the distribution for middle-aged men is approximately symmetric. ✓

(E) The distribution for young adult men is skewed to the left, and the distribution for middle-aged men is skewed to the right.

36. The yields, in bushels per acre, of soybeans for 25 randomly selected counties in the midwestern region of the country in 2016 are summarized in the table.

n	$\bar{x}$	s
25	45.88	7.94

Assuming all conditions for inference are met, which of the following is a 99 percent confidence interval for the population mean yield of soybeans for all counties in the midwestern region in 2016?

(A) $45.88 \pm 2.797\left(\frac{7.94}{25}\right)$

(B) $45.88 \pm 2.797\left(\frac{7.94}{5}\right)$ ✓

(C) $45.88 \pm 2.576\left(\frac{7.94}{5}\right)$

(D) $45.88 \pm 2.787\sqrt{\frac{7.94}{25}}$

(E) $45.88 \pm 2.797\sqrt{\frac{7.94}{25}}$

37. Ming is learning the game of tennis and will practice serving a tennis ball within bounds. Assume that the probability of Ming serving a tennis ball within bounds is 0.40 and that her serves are independent of each other.

What is the probability that the first time Ming serves a tennis ball within bounds will occur on her 4th attempt?

(A) 0.4

(B) $1-(0.4)^4$

(C) $(0.6)^3\left(0.4\right)$ ✓

(D) $\binom{4}{1}(0.4)(0.6)^3$

(E) $\binom{4}{1}(0.4)(0.6)^3+\binom{4}{2}(0.4)^2(0.6)^2+\binom{4}{3}(0.4)^3(0.6)$

2022 Form I

38. Caleb designed an experiment to investigate whether listening to music or chewing gum affects the time to complete a level of a certain video game. Before the game started, each person in a sample was randomly assigned to either listen to music or not listen to music. Each person was also randomly assigned to either chew gum or not chew gum. At the end of the game, the time to complete the level for each person in the sample was recorded.

Which of the following correctly identifies the response variable of the experiment and whether the variable is categorical or quantitative?

(A) The response variable is the time to complete the level and the variable is quantitative. ✓

(B) The response variable is the time to complete the level and the variable is categorical.

(C) The response variable is whether the person was assigned to listen to music or chew gum and the variable is categorical.

(D) The response variable is whether listening to music or chewing gum affects time to complete the level and the variable is quantitative.

(E) The response variable is whether listening to music or chewing gum affects time to complete the level and the variable is categorical.

39. For a particular video game, Ariana's mean number of points earned per game is 190, and the standard deviation is 40. Suppose that she plays 50 games in one evening, and her goal is to obtain a mean score of at least 200 for the 50 games.

How many standard deviations above 190 is her goal?

(A) 0.25

(B) 1.41

(C) 1.77 ✓

(D) 11.18

(E) 12.50

40. The distribution of weight of a small watermelon is known to have a mean of 5 pounds and a standard deviation of 1.2 pounds. Six small watermelons fit in a crate, and the crate weighs exactly 1 pound.

Assuming the weights of small watermelons are independent, what is the standard deviation of the total weight of a random sample of 6 small watermelons and the crate?

(A) $\sqrt{6^2(1.2)^2}$

(B) $1+\sqrt{6^2(1.2)^2}$

(C) $1+6^2(1.2)^2$

(D) $\sqrt{6(1.2)^2}$ ✓

(E) $1+\sqrt{6(1.2)^2}$

2022 Form I

41. Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

A random sample of 25 cans of soda was selected from a factory that produces soda. The amount of soda, in milliliters (ml), contained in each sampled can was measured. The data are shown in the following stem-and-leaf plot.

34	7
35	1 2 3 3 4 4
35	5 5 5 6 7 9
36	2
36	5 6 6 7 8 8
37	0 3 3 4
37	5

Key: 34|7 represents 347 ml of soda

(a) Describe the distribution of amount of soda in a can.

The factory uses two machines to fill the soda cans: Machine A and Machine B. The following back-to-back stem-and-leaf plot shows the soda amounts from the same random sample categorized by machine.

Machine A							Machine B					
						34	7					
					4	35	1	2	3	3	4	
					5	35	5	5	6	7	9	
					2	36						
8	8	7	6	6	5	36						
		4	3	3	0	37						
					5	37						

Key: 34|7 represents 347 ml of soda

(b) Calculate the median and IQR for each machine.

2022 Form I

Machine A median: Machine B median:

Machine A IQR: Machine B IQR:

(c) Compare the variability of the measurements in the amount of soda filled per can by the two machines by referencing the appropriate measure calculated in part (b). Justify your answer.

Part A, Part B, Part C

Select a point value to view scoring criteria, solutions, and/or examples and to score the response

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4.
- The model solution represents an ideal response to each part of the question and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Scoring

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

0	1	2	3	4 ✓

All three parts essentially correct.

- [ ] Part (a) essentially correct
- [ ] Part (a) partially correct
- [ ] Part (a) incorrect
- [ ] Part (b) essentially correct
- [ ] Part (b) partially correct
- [ ] Part (b) incorrect
- [ ] Part (c) essentially correct
- [ ] Part (c) partially correct
- [ ] Part (c) incorrect

Solution

Part (a):

2022 Form I

The distribution of amount of soda in a can is bimodal and approximately symmetric. The center is around 362 ml and all amounts are between 347 ml and 375 ml. There are no apparent outliers.

Scoring

Part (a) is scored as follows:

Essentially correct (E) if the description of the distribution satisfies at least three of components 1-4 and component 5:

Shape: The distribution is bimodal

Center: Refers to at least one of the following: the center of the distribution is close to 362 ml, the median is 359 ml, or the mean is 361 ml

Spread: Refers to at least one of the following: the range is $375 - 347 = 28$ ml, all values are between 347 ml and 375 ml, or the IQR is $368 - 354 = 14$ ml

Unusual features: Refers to at least one of the following: no outliers, no gaps, or no unusual features

Includes context of amount of soda or milliliters (ml)

Partially correct (P) if the response satisfies any three of components out of components 1-5.

Incorrect (I) if the response does not meet the criteria for E or P.
Additional Notes:

- A response that interprets the key of the stem-and-leaf plot incorrectly (for example, interprets $34|7$ as "34.7 ml ") reduces the score from an E to a P or P to an I.
- Shape:
    - Component 1 cannot be satisfied if a response describes the stem-and-leaf plot as unimodal or normal or approximately normal.
    - A response that only refers to a gap does not satisfy component 1.
    - A response that refers to a “dip” near 362 satisfies the bimodal shape component.
    - A response that addresses the center or spread of each of the two clusters satisfies component 1.
    - A response that addresses only symmetry, while appropriate, does not satisfy component 1.
- Center:
    - A response that addresses center using interval language such as “the mean of the distribution is between 359 and 365” must, for any single measure of center, provide an interval with a lower endpoint no less than 359 ml, and an upper endpoint no greater than 365 ml to satisfy component 2.
    - A response that addresses center using approximate language such as “the mean of the distribution is approximately 362” must, for any single measure of center, specify a numeric value that is not less than 359 ml and that is not greater than 365 ml to satisfy component 2.
    - A response that uses definitive language such as “the mean of the distribution is 361.1” must identify the corresponding numeric value correctly to satisfy component 2.

2022 Form I

- The median of the distribution is 359 ml.
- The mean of the distribution is 361.1 (or 361.12) ml .
- The center (or average) of the distribution can be the value of the correct median or mean.

- Spread:
  - A response recognizing all values in the sample fall between 347 and 375 ml satisfies component 3 only for these exact endpoints and need not appeal to a specific measure of spread such as the range.
  - A response that uses interval language must use it correctly. For example, "the observations range from 347 to 375" satisfies component 3 because it correctly indicates that all observations are between 347 and 375, inclusive. However, a statement such as "the range is between 347 and 375" is incorrect because the range is a single number, i.e., $375 - 347 = 28$ ml.
  - A response that appeals to a specific measure of spread using approximate language, such as "the IQR is approximately 14," must specify a numeric value within the bounds appropriate to that measure of spread. For range, the value must be 28 ml, for IQR, the value must be between 10 and 18 ml (inclusive); for standard deviation, the value must be between 8.0 and 8.5 ml (inclusive).
  - A response that appeals to a specific measure of spread using definitive language, such as "the range of the distribution is 28," must identify the corresponding numeric value correctly to satisfy component 3.
    - The range of the distribution is 28 ml.
    - The IQR of the distribution is 14 ml.
    - The standard deviation of the distribution is 8.393 (or 8.4) ml.
    - The spread of the distribution can be the value of the correct range, IQR, or standard deviation.
- Unusual Features:
  - Component 4 cannot be satisfied if a response indicates that the distribution has an unusual feature.

Solution

Part (b):

The median amount of soda in the cans filled by Machine A is 367.5 ml and the median amount of soda in the cans filled by Machine B is 354 ml.

The IQR of Machine A is $Q_3 - Q_1 = 373 - 365 = 8$ ml and the IQR of Machine B is $Q_3 - Q_1 = 356 - 352 = 4$ ml.

Scoring

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

The median is correct for Machine A

2022 Form I

The median is correct for Machine B

The IQR is correct for Machine A

The IQR is correct for Machine B

Partially correct (P) if the response satisfies two or three of the four components.
Incorrect (I) if the response does not meet the criteria for E or P.
Additional Notes:

- A response that interprets the key of the stem-and-leaf plot incorrectly (for example interprets $34|7$ as "34.7 ml ") reduces the score from E to P or P to I. However, if the score in part (a) was reduced for this error, part (b) may still be scored E.

- For machine A, any median value between 367 and 368, inclusive, may satisfy component 1.

- Different methods for computing quartiles will result in slightly different values for the IQRs for the two machines, but the IQR for machine B should be smaller than the IQR for machine A. The IQR for machine B should be scored correct if the value is between 3 and 4.5, inclusive and the IQR for machine A should be scored correct if the value is between 8 and 9.5, inclusive.

<u>Solution</u>

Part (c):

Machine B appears to have less variability than Machine A because the IQR of Machine B, which is 4 ml per can, is smaller than the IQR of Machine A, which is 8 ml per can.

<u>Scoring</u>

Part (c) is scored as follows:

Essentially correct (E) if the response compares the IQR values calculated in part (b) for the two machines and concludes that the amounts of soda per can are less variable for the machine with the smaller value for the IQR.

Partially correct (P) if the response compares the IQR values computed in part (b), but does not provide a conclusion about which machine dispenses less variable amounts of soda

OR

if the response provides a correct conclusion (i.e., states Machine A is more variable, or Machine B is less variable) and states the IQRs, but never compares the IQRs

OR

if the response correctly compares measures of variability other than the IQR (e.g., compares ranges or compares standard deviations) and states a correct conclusion

OR

if the response indicates that amounts of soda are less variable for Machine B with justification that the data are less spread out for Machine B than for Machine A, without comparing the IQR values for the two machines.

2022 Form I

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

- A response that includes a comparison of medians, as part of the justification for comparing variability, should be considered a parallel solution and be scored as incorrect.
- A response that does not calculate the values of the IQRs in part (b) may still be scored E if the IQRs are correctly calculated and compared in part (c).

42. Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

An educational researcher has noticed that because many students can type faster than they can write, lecture notes taken by students using laptops tend to look like a transcript of the instructor's lecture, while handwritten lecture notes tend to be shorter summaries of the key points made during the lecture. The researcher wants to investigate whether the difference in note-taking methods has an impact on student quiz performance and is considering the following two proposals for the investigation.

- Proposal I: The researcher will select a random sample of students from a list of volunteers and allow the students to use their preferred note-taking method, laptops or handwritten.
- Proposal II: The researcher will randomly assign the note-taking method (laptops or handwritten) to each student in a group of volunteers, so that half of the volunteers are assigned to each method.

(a) Which of the two proposals is not an experiment? Justify your answer.

(b) Could either of the two proposals be used to establish a cause-and-effect relationship? Justify your answer.

(c) Describe how student study habits could be a confounding variable when using Proposal I for the investigation.

Part A, Part B, Part C

Select a point value to view scoring criteria, solutions, and/or examples and to score the response

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4.
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Scoring

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

2022 Form I

0	1	2	3	4 ✓

All three parts essentially correct.

- [ ] Part (a) essentially correct
- [ ] Part (a) partially correct
- [ ] Part (a) incorrect
- [ ] Part (b) essentially correct
- [ ] Part (b) partially correct
- [ ] Part (b) incorrect
- [ ] Part (c) essentially correct
- [ ] Part (c) partially correct
- [ ] Part (c) incorrect

Solution

Part (a):

Proposal I is an observational study, not an experiment, because the two note-taking methods will not be assigned to the students by the researchers. Each student will make their own decision about whether to use a laptop or take hand-written notes.

Proposal II is an experiment because the two note-taking methods will be assigned to the volunteers by the researcher.

Scoring

Part (a) is scored as follows:

Essentially correct (E) if the response indicates that Proposal I is not an experiment and provides an explanation based on the fact that the note-taking method used by each student in the study will not be determined by the researchers

OR

that indicates each student will self-select a note-taking method

OR

indicates that note-taking methods will be ONLY observed

Partially correct (P) if the response indicates that Proposal I is not an experiment and provides an explanation indicating that treatments will not be assigned by the researchers or students will self-select a treatment or there are no treatments but does not include context (note-taking methods)

OR

the response indicates that Proposal I is an observational study but does not provide an explanation indicating that researchers will not assign the note-taking method used by each student in the study

OR

the response indicates that Proposal I is not an experiment and provides an explanation that states that note-taking methods will be observed without a modifier for observed (such as only, just, merely, or simply) and without indicating that treatments will not be assigned by the researchers.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

- If the response includes one or more incorrect criteria for distinguishing between an experiment and an observational study (such as "it is an observational study because an experiment must have a control group"), the score is lowered by one level from E to P or from P to I.
- A response that indicates that Proposal II is not an experiment, or both Proposal I and Proposal II are not experiments, should be scored I.
- Any indication that Proposal II is an experiment should be considered extraneous.

Solution

Part (b):

Yes, Proposal II can be used to establish a cause-and-effect relationship, because the researcher will randomly assign the treatments (note-taking methods) to the subjects (members of a group of volunteers).

Proposal I cannot be used to establish a cause-and-effect relationship because note-taking methods will not be randomly assigned to the volunteers.

Scoring

Part (b) is scored as follows:

Essentially correct (E) if the response selects Proposal II and provides a justification based on the random assignment of treatments to subjects (or that Proposal II is a randomized experiment).

Partially correct (P) if the response selects Proposal II and only indicates that it is an experiment or indicates treatments will be assigned but does not include "randomly".

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

- A response that selects Proposal I, or selects both Proposal I and Proposal II, should be scored I.

Solution

Part (c):

Student study habits would be a confounding variable if a greater proportion of students with weaker study habits use laptops to take notes and a greater proportion of students with stronger study habits take notes by hand, and regardless of

2022 Form I

the note-taking method they use, students with weaker study habits tend to have lower quiz scores and students with stronger study habits tend to have higher quiz scores. This would result in lower test scores for the students using laptops and higher test scores for students taking notes by hand, even when the note-taking method does not have an effect on quiz performance.

Scoring

Part (c) is scored as follows:

Essentially correct (E) if the response provides an explanation of how student study habits could be a confounding variable that includes the following two components:

Indicates that there is an association between student study habits and quiz performance

Indicates that there is an association between student study habits and note-taking methods

Partially correct (P) if the explanation satisfies only one of the two components

OR

if the explanation includes both components but one or both components lack context. For example, component 1 indicates that there is an association between student study habits and the response, without mentioning performance on quizzes, or component 2 indicates that there is an association between student study habits and the explanatory variable without mentioning note-taking methods.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

- The term “correlation” may not be used in place of “association” in component 1, because student study habits cannot be recorded as a quantitative variable.
- The term “correlation” may not be used in place of “association” in component 2, because note-taking method cannot be recorded as a quantitative variable.

2022 Form I

43. Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

In the western United States, water is sometimes diverted from mountain streams and rivers to be used for agriculture and municipal water supplies. Diverting too much water raises the water temperature in the mountain streams and rivers, affecting the fish that live there.

A certain mountain river is abundant in brown trout and has water diverted from it for city use. Assume that the water temperature along all locations of the river has a distribution with mean $18^\circ\mathrm{C}$ and standard deviation of $2^\circ\mathrm{C}$.

A sport-fishing group plans to measure the water temperature at 64 randomly selected locations along the river. Let the random variable F represent the mean water temperature, in degrees Celsius, of 64 locations along the river selected at random by the sport-fishing group.

(a) Describe the distribution of F.

(b) Fish, such as brown trout, become stressed once water temperatures exceed $18.3^\circ\mathrm{C}$. Determine the probability F will exceed $18.3^\circ\mathrm{C}$. Show your work.

A large metropolitan water district that uses water diverted from the same river is also planning to measure the water temperature at 100 randomly selected locations along the river. Because of sampling variability, the sample collected by the water district is likely to yield a sample mean water temperature different from the sample mean water temperature collected by the sport-fishing group. Let the random variable D represent the mean water temperature, in degrees Celsius, of 100 locations along the river selected at random by the large metropolitan water district.

(c) Describe the distribution of $D - F$.

Part A, Part B, Part C

Select a point value to view scoring criteria, solutions, and/or examples and to score the response

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4.
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

<u>Scoring</u>

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

0	1	2	3	4 ✓

2022 Form I

All three parts essentially correct.

- ☐ Part (a) essentially correct
- ☐ Part (a) partially correct
- ☐ Part (a) incorrect
- ☐ Part (b) essentially correct
- ☐ Part (b) partially correct
- ☐ Part (b) incorrect
- ☐ Part (c) essentially correct
- ☐ Part (c) partially correct
- ☐ Part (c) incorrect

Solution

Part (a):

The distribution of random variable *F* will have a mean of 18°C and a standard deviation of $\frac{2}{\sqrt{64}} = 0.25^\circ\text{C}$. The shape will be approximately normal because the sample size $(n = 64)$ is large enough, $n > 30$.

Scoring

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

The correct mean (18°C) is given

The correct standard deviation (0.25°C) is given with supporting work shown

The shape of the distribution is described as approximately normal

Partially correct (P) if the response satisfies only two of the three components

OR

if the response only satisfies component 2.

Incorrect (I) if the response does not satisfy the criteria for E or P.
Additional Notes:

Component 1

- Phrases such as “the center” or “centered around” are minimally acceptable ways to identify the mean for component 1.

Component 2

- A standard deviation of $\frac{2}{\sqrt{64}}$ satisfies component 2. It is not necessary to simplify the standard deviation to

2022 Form I

satisfy component 2.

- To satisfy the supporting work criterion of component 2 the response must clearly indicate that the standard deviation for F is computed by dividing 2 by the square root of 64 (or dividing by 8). This may be indicated with words, $\frac{2}{\sqrt{64}}$, $\frac{2}{8}$, or standard notation such as $\frac{\sigma}{\sqrt{n}}$.

- If a response indicates 0.25 without supporting work in part (a), component 2 may be satisfied if supporting work is shown in part (b).

Component 3

- A response may satisfy component 3 by any of the following:
  - Graphical: Displaying a graph of a normal density function.
  - Words: Using a statement such as “approximate normal distribution.”
  - Standard notation: Notations such as $F \sim N\left(18, \frac{2}{\sqrt{64}}\right)$ or $F \sim N\left(18, \left(\frac{2}{\sqrt{64}}\right)^2\right)$ satisfy all three components, but the notation $F \sim N(18, 0.25)$ satisfies only components 1 and 3 because it does not show how the standard deviation was calculated. Neither the correct mean nor the correct standard deviation must be reported to satisfy component 3.

- The response does not need to give a justification for shape to satisfy component 3. However, if a justification is given, it must be correct to satisfy component 3.

- Indication of a “normal distribution” instead of an “approximately normal distribution” should be viewed as a minor error that does not reduce the score.

General

- For part (a), errors in notation such as using $s_{\bar{x}}$ instead of σ_F or using $\mu_{\bar{x}}$ instead of μ_F should be viewed as minor errors that do not reduce the score.

- An arithmetic or transcription error in a response can be ignored if correct work is shown.

Solution

Part (b):

$$P(F > 18.3) = P\left(Z > \frac{18.3 - 18.0}{\frac{2}{\sqrt{64}}}\right)$$
$$= P(Z > 1.2) \approx 0.1151$$

Scoring

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

Indicates the use of a normal (or approximately normal) distribution and identifies the correct parameter values (mean 18

2022 Form I

and standard deviation $\frac{2}{\sqrt{64}}$ or 0.25) or values reported in part (a)

Specifies the correct event (boundary value and direction) or an event consistent with values reported in component 1

Provides the correct probability of 0.1151 or the probability consistent with components 1 and 2

Partially correct (P) if the response satisfies only two of the three components

OR

if the response fails to satisfy component 1 and 2, but shows the correct *z*-score formula, *z*-score value, and correct probability. (e.g., $\frac{18.3-18.0}{0.25} = 1.2$, resulting in a probability of 0.1151).

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

Component 1

- A response may satisfy component 1 by any of the following or a combination of the following:
    - Graphical: Displaying a graph of a normal density function with the appropriate scale on the horizontal axis showing the mean and standard deviation for the distribution of sample mean water temperature.
    - Calculator function syntax: Labeling correct values of the mean and standard deviation in a "normalcdf" statement, such as $\text{normalcdf}\left(\text{lower} = 18.3,\ \text{upper} = \infty, \text{mean} = 18, \text{standard deviation} = \frac{2}{\sqrt{64}}\right)$. Correct specification of the lower and upper bounds is not required to satisfy component 1.
    - Words: Using a statement such as "approximately normal distribution with mean 18 and standard deviation $\frac{2}{\sqrt{64}}$."
    - Standard notation: Using standard notation such as $N(18, 0.25)$ or $N\left(18, (0.25)^2\right)$
    - *Z*-score: Displaying the correct mean and standard deviation in a *z*-score calculation that includes "*z*," such as $z = \frac{18.3-18}{0.25}$.
- If no value for the mean is reported in part (a), then the correct value (18) must be indicated in part (b) to satisfy component 1.
- If no value for the standard deviation is reported in part (a), then either 0.25 or 2 may be used in part (b) to satisfy component 1.

Component 2

- A response may satisfy component 2 by any of the following or a combination of the following:
    - Graphical: Displaying a graph of a normal density function with the region of interest ($F > 18.3$ or $Z > 1.2$) clearly identified. The shaded area does not need to be proportional, but the boundary should be on the proper side of the mean and the shading should be in the proper direction.

- Calculator function syntax: Labeling the lower and upper bounds of the region of interest in a "normalcdf" statement, such as:

  $\text{normalcdf}(\text{lower} = 18.3, \text{upper} = \infty, \text{mean} = 18, \text{standard deviation} = 0.25).$

  $\text{normalcdf}(\text{lower} = 1.2, \text{upper} = \infty, \mu = 0, \sigma = 1).$

- Words: Specifying the correct event in words with the correct numerical value for the boundary value and correct direction such as "The probability that *F* will exceed 18.3°C …" or $P(\text{mean water temperature} > 18.3).$

- Standard notation: Using standard notation such as $P(F > 18.3)$ or $P\left(Z > \frac{18.3-18}{0.25}\right)$ or $P(Z > 1.2).$

- A response that fails to satisfy component 1 and 2, but shows a *z*-score formula, *z*-score value, and probability consistent with using a standard deviation of 2 (e.g., $\frac{18.3-18.0}{2} = 0.15$, resulting in a probability of 0.4404) should be scored P.

General

- It is not necessary to define the random variable *F* because it is defined in the stem. It is not necessary to define the random variable *Z* because it is standard notation. Any other random variable must be defined correctly.

- An error in statistical notation, such as using *s* instead of σ for the population standard deviation or using $\overline{F}$ instead of μ for the population mean, does not satisfy component 1.

- If the only error in the response to part (b) is the reversal of the numerator for the *z*-score $(18 - 18.3)$, the response is scored P.

- An arithmetic or transcription error in a response can be ignored if correct work is shown.

Solution

Part (c):

The sampling distribution of $D - F$ will have a mean of $18^\circ\text{C} - 18^\circ\text{C} = 0^\circ\text{C}$ and a standard deviation of $\sqrt{\frac{2^2}{64} + \frac{2^2}{100}} = \sqrt{0.1025} \approx 0.32^\circ\text{C}.$ The shape will be approximately normal because the sample sizes are large enough, both greater than 30.

Scoring

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

The correct value (0), or a value consistent with the response to part (a), of the mean of the sampling distribution for the difference in the two sample means is given

The correct value (0.32), or a value consistent with the response to part (a), of the standard deviation of the sampling distribution for the difference in the two sample means is given with supporting work shown

The shape of the distribution is described as approximately normal

2022 Form I

Partially correct (P) if the solution satisfies only two of the three components

OR

satisfies only component 2.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

Component 1

- Phrases such as “the center” or “centered around” are minimally acceptable ways to identify the mean.

Component 2

- To satisfy the supporting work criterion of component 2 the response must clearly indicate that the standard deviation for $D - F$ is computed as the square root of the sum of the variances of *D* and *F* (e.g., $\sqrt{\frac{2^2}{64} + \frac{2^2}{100}}$ or $\sqrt{0.25^2 + 0.2^2}$). It is not necessary to simplify this formula or report the value as 0.32 to satisfy component 2.

Component 3

- A response may satisfy component 3 by any of the following:
  - Graphical: Displaying a graph of a normal density function.
  - Words: Using a statement such as “approximate normal distribution.”
  - Standard notation: Notations such as $D - F \sim N\left(0, \sqrt{\frac{2^2}{64} + \frac{2^2}{100}}\right)$ or $D - F \sim N\left(0, \left(\sqrt{\frac{2^2}{64} + \frac{2^2}{100}}\right)^2\right)$ satisfy all three components, but the notation $D - F \sim N(0, 0.32)$ satisfies only components 1 and 3 because it does not show how the standard deviation was calculated. Neither the correct mean nor the correct standard deviation for $D - F$ must be reported to satisfy component 3.
- The response does not need to give a justification for shape to satisfy component 3. However, if a justification is given, it must be correct to satisfy component 3.
- Indication of a “normal distribution” instead of an “approximately normal distribution” should be viewed as a minor error that does not reduce the score.

General

- For part (c), errors in notation such as using s_{D-F} instead of σ_{D-F} or using $\bar{x}_D - \bar{x}_F$ instead of $\mu_D - \mu_F$ should be viewed as a minor error that does not reduce the score.
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

2022 Form I

44. Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

A veterinarian is comparing two diets, Diet J and Diet W, designed to help overweight large-breed dogs lose weight. Thirty-two overweight large-breed dogs were randomly assigned to one of the two diets, with 16 dogs in each group. The dogs remained on the assigned diet for six months and the weight loss, in pounds, was recorded for each dog. Summary statistics of the weight loss for the two groups are shown.

	Mean (pounds)	**Standard Deviation (pounds)**
Diet J	7.523	0.819
Diet W	6.379	1.012

Assume the distribution of weight loss for all overweight large-breed dogs who use Diet J is approximately normal and assume the distribution of weight loss for all overweight large-breed dogs who use Diet W is approximately normal. Do the data provide convincing statistical evidence, at the significance level of $\alpha = 0.05$, that Diet J is more effective than Diet W in terms of mean weight loss for overweight large-breed dogs that are similar to those used in the study? Complete the appropriate inference procedure to support your answer.

3-section Inference Scoring

Select a point value to view scoring criteria, solutions, and/or examples and to score the response

- This question is scored in three sections. Each section is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The first section includes statements of the null and alternative hypotheses and identification of the appropriate hypothesis test. The second section includes verifying the conditions for the identified test and calculating the value of the test statistic and the corresponding *p*-value. The third section includes the conclusion for test identified test. The response is then categorized based on the scores assigned to each section and awarded an integer score between 0 and 4.

- The model solution represents an ideal response to each section of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Scoring

Scoring sections 1, 2, and 3 are each scored as essentially correct (E), partially correct (P), or incorrect (I).

0	1	2	3	4

All three sections essentially correct.

- [ ] Scoring Section 1 essentially correct

2022 Form I

- [ ] Scoring Section 1 partially correct
- [ ] Scoring Section 1 incorrect
- [ ] Scoring Section 2 essentially correct
- [ ] Scoring Section 2 partially correct
- [ ] Scoring Section 2 incorrect
- [ ] Scoring Section 3 essentially correct
- [ ] Scoring Section 3 partially correct
- [ ] Scoring Section 3 incorrect

Solution

Section 1:

Let μ_J represent the true mean weight loss of all overweight large-breed dogs who would be fed Diet J and let μ_W represent the true mean weight loss of all overweight large-breed dogs who would be fed Diet W.

The null and alternative hypotheses are

$H_0 : \mu_J = \mu_W$
$H_a : \mu_J > \mu_W$

An appropriate inference procedure is a two-sample *t*-test for a difference of two population means.

Scoring

Section 1 is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

States a correct null hypothesis of equal means AND a correct one-sided alternative hypothesis

Provides sufficient context by including references to:

- the two groups (types of diets; Diet J and Diet W)
- *populations* (all overweight large-breed dogs similar to those in the study)
- the response variable (weight loss)

Identifies a two-sample *t*-test for a difference in population means by name
(e.g., "two-sample *t*-test") or by formula

Partially correct (P) if the response does not meet the criteria for E but satisfies either component 1 and/or component 3.

Incorrect (I) if the response does not meet the criteria for E or P.
Additional Notes:

- The elements of component 2 do not have to be satisfied with the statement of the hypotheses. They may be satisfied by work presented anywhere in the response, most likely by the statement of the conclusion.

2022 Form I

- If the statement of the hypotheses refers to population means and the conclusion refers to sample means (or vice versa), then the population aspect of component 2 is not satisfied.

- A response that states the hypotheses in words (e.g., "the null hypothesis is that the means are equal, and the alternative hypothesis is that the mean of Diet J is greater than the mean of Diet W") may satisfy component 1. Neither context nor the concept of *population* is required to satisfy component 1.

- To satisfy component 1, the hypotheses must be stated in terms of two means. Use of standard symbols for means with labels that indicate two different groups (such as $\mu_{\text{Diet J}}$ and $\mu_{\text{Diet W}}$, or $\bar{x}_J$ and $\bar{x}_W$) is sufficient to satisfy component 1. Any other symbol must be clearly defined as a mean. Labels such as "J" and "W" must be included and the correct alternative hypothesis must be identified for the response to satisfy component 1.

- The notation μ_J and μ_W satisfies the group and population aspects of component 2, because it is standard notation for population means and it reflects the context of the two types of diets.

- Without additional description that provides links to Diet J and Diet W, notation such as μ_1 and μ_2 does not satisfy either the group aspect of component 2 or the identification of the correct alternative hypothesis for component 1.

- A response that states the hypotheses in words may satisfy some or all aspects of component 2. For example, "the null hypothesis is that the mean weight loss of all large-breed dogs who would be fed Diet J is equal to the mean weight loss of all large-breed dogs fed Diet W and the alternative hypothesis is that the mean weight loss of all large-breed dogs who would be fed Diet J is greater than the mean weight loss of all large-breed dogs fed Diet W" satisfies all aspects of component 2.

- If the response clearly refers to *sample* means instead of *population* means using words or symbols (e.g., $\bar{x}_J$ and $\bar{x}_W$, or $\bar{x}_{\text{Diet J}}$ and $\bar{x}_{\text{Diet W}}$), then the population aspect of component 2 is not satisfied unless the symbols used are defined as *population* means.

- If the response identifies the correct test by name, but also states an incorrect formula or uses incorrect notation in the formula, then component 3 is not satisfied.

- If the response refers to a *t*-test for "mean difference" (instead of "difference in means"), then component 3 is not satisfied.

- A pooled two-sample *t*-test may satisfy component 3 if the response comments on the plausibility of the equal population variances assumption.

- If the response identifies a two-sample *z*-test for a difference in population means, then component 3 is not satisfied.

Matched-Pairs Approach*:*

- Because a response presenting a matched-pairs *t*-test approach cannot satisfy component 3, such a response cannot meet the criteria for Essentially Correct, E. If the response satisfies component 1, it may be scored P. Otherwise, it is scored I. (Note that null and alternative hypotheses about (μ-dif) satisfy component 1 ONLY if matched pairs is explicitly named as the approach.)

- A response that attempts a matched-pairs approach using a confidence interval may be scored P if component 2 is satisfied and the response indicates that the confidence interval is for the mean difference. Otherwise, it is scored I.

Confidence Interval Approach*:*

2022 Form I

- For a response using a confidence interval approach, component 1 is satisfied if the response clearly indicates that the confidence interval is used to assess the correct one-sided alternative to the null hypothesis of equal means.

- A response that uses a two-sample *t*-interval for a difference in population means must demonstrate all three aspects (groups, population, and response variable) to satisfy component 2. For example, component 2 is satisfied if the response indicates that it is a confidence interval for the difference between the mean weight loss of *all* overweight large breed-dogs who would be fed diet J and the mean weight loss of *all* overweight large breed-dogs who would be fed diet W.

- If a two-sample *t*-interval for a difference in population means is identified correctly by name or by formula, then component 3 is satisfied.

- If the response uses individual one-sample *t*-intervals for the two means, which is not a correct approach, then component 3 is not satisfied.

Solution

Section 2:

The independent observations condition for performing the two-sample *t*-test for a difference in population means is satisfied because the data were obtained from a randomized experiment where the 32 overweight large-breed dogs were randomly assigned to the two diets.

Since the distribution of weight loss for all overweight large-breed dogs on each diet is approximately normal, then the sampling distribution of $\bar{x}_J - \bar{x}_W$ is also approximately normal.

The value of the *t*-statistic is

$$t = \frac{\bar{x}_J - \bar{x}_W}{\sqrt{\frac{s_J^2}{n_J} + \frac{s_W^2}{n_W}}} = \frac{7.523 - 6.379}{\sqrt{\frac{(0.819)^2}{16} + \frac{(1.012)^2}{16}}}$$

$$\approx \frac{1.144}{0.325} \approx 3.515$$

The corresponding *p*-value, using an approximate degrees of freedom of 28.75, is $P(t > 3.515) \approx 0.0007.$

Scoring

Section 2 is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

Checks the independence condition by referring to the randomized experiment

Indicates that the approximately normal distribution of weight loss of all overweight large-breed dogs for BOTH diets supports the assumption that the sampling distribution of $\bar{x}_J - \bar{x}_W$ is approximately normal

Correctly reports the value of the *t*-statistic consistent with the stated alternative hypothesis

Correctly reports the *p*-value, consistent with the stated alternative hypothesis

Partially correct (P) if the response satisfies only two or three of the four components.
Incorrect (I) if the response does not meet the criteria for E or P.

2022 Form I

Additional Notes:

- Component 1 is not satisfied if the response indicates that the independent condition is met because the sample was selected at random from all overweight large-breed dogs.
- Component 1 is satisfied if the response indicates that dogs were randomly assigned to groups or diets, even if the terms ‘independent’ or ‘independence’ are not used.
- In order to satisfy the check on approximately normal distribution of weight loss of all overweight large-breed dogs using both diets, it is minimally acceptable to state “normal in BOTH diets.” However, component 2 is not satisfied if the response implies that weight loss is approximately normally distributed for only one diet.
- Values for the test statistic and p-value depend on how the degrees of freedom were determined. The following combinations of t-values and p-values all satisfy components 3 and 4, respectively:
    - The degrees of freedom for $t \approx 3.515$ from the model solution were computed with the Satterthwaite approximation used by many calculators:

$$df = \frac{\left(\frac{s_J^2}{n_J} + \frac{s_W^2}{n_W}\right)^2}{\frac{1}{n_J - 1}\left(\frac{s_J^2}{n_J}\right)^2 + \frac{1}{n_W - 1}\left(\frac{s_W^2}{n_W}\right)^2} = \frac{\left(\frac{(0.819)^2}{16} + \frac{(1.012)^2}{16}\right)^2}{\frac{1}{16-1}\left(\frac{(0.819)^2}{16}\right)^2 + \frac{1}{16-1}\left(\frac{(1.012)^2}{16}\right)^2}$$
$$\approx 28.75$$

    - Using $t \approx 3.515$ and $(16 - 1) = 15$ degrees of freedom, the resulting p-value is 0.0016.
    - Using a pooled t-test with $t \approx 3.515$ and 30 degrees of freedom (which would be appropriate because the standard errors for the two samples are similar), the pooled estimate of the standard error is 0.92 and the resulting p-value is 0.0007.
- Discussion of degrees of freedom will be treated as extraneous in scoring for Section 2.
- If the response incorrectly identifies the test as a z-test in section 1, the correct z-statistic of 3.515 satisfies component 3 and a p-value of 0.0002 satisfies component 4.
- If the response does not identify the z-test in section 1 but reports the test statistic as $z = 3.515$ (instead of t) then component 3 is not satisfied; however, component 4 may still be satisfied if the correct p-value of 0.0002 is reported with a z-statistic of 3.515.
- If the response compares the value of the test statistic to a critical value instead of computing a p-value, then a comparison consistent with the stated alternative hypothesis, satisfies component 4.
- If a two-tailed alternative hypothesis is stated, then the p-value must be consistent with the stated alternative hypothesis to satisfy component 4.
- A response that reports the correct value for the t-statistic but contains errors in supporting work may still satisfy component 3.
- If the response satisfies component 4, any supporting work for the p-value may be treated as extraneous.

Matched-Pairs Approach:

- Section 2 of a response using a matched-pairs t-test or a matched-pairs confidence interval must be scored I.

2022 Form I

Confidence Interval Approach:

- If the stated alternative hypothesis is correct or no alternative hypothesis is provided:
    - If either a one-sided 95 percent confidence interval for $\mu_{\text{Diet J}} - \mu_{\text{Diet W}}$ is correctly calculated as $(0.590, \infty)$ or a two-sided 90 percent confidence interval for $\mu_{\text{Diet J}} - \mu_{\text{Diet W}}$ is correctly calculated as $(0.590, 1.698)$ then component 3 is satisfied.
    - If only the lower end of the confidence interval for $\mu_{\text{Diet J}} - \mu_{\text{Diet W}}$ is used to reach a conclusion, then component 4 is satisfied.
- If the stated alternative hypothesis is incorrect (2-sided or reversed direction), the confidence interval approach must be consistent with the stated alternative to satisfy components 3 and 4:
    - An interval consistent with the stated alternative will satisfy component 3. A two-sided 95 percent confidence interval for $\mu_{\text{Diet J}} - \mu_{\text{Diet W}}$ is $(0.479, 1.808)$ and a lower one-sided 95 percent confidence interval for $\mu_{\text{Diet J}} - \mu_{\text{Diet W}}$ is $(-\infty, 1.808)$.
    - If the two-sided confidence interval is correctly interpreted based on whether zero is in the interval, then component 4 is satisfied; if only the upper end of the lower one-sided confidence interval is used to reach a conclusion, then component 4 is satisfied.
- $\text{Diet W} - \text{Diet J}$ intervals can be used as well as they will be the negative of the intervals given.

<u>Solution</u>

Section 3:

Because the *p*-value of 0.0007 is less than the significance level of 0.05, the null hypothesis should be rejected. The data provide convincing statistical evidence that for dogs similar to those used in the study that the mean weight loss for overweight large-breed dogs fed Diet J is greater than the mean weight loss for overweight large-breed dogs fed Diet W.

<u>Scoring</u>

Section 3 is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

Provides a correct justification of the conclusion based on whether the *p*-value is less than $\alpha = 0.05$ (or a comparison of the value of the test statistic to an appropriate critical value)

States a correct conclusion consistent with the stated alternative hypothesis OR states a correct conclusion that answers the inference question

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not meet the criteria for E or P.
Additional Notes:

- Although including proper context (the concept of population proportion and referencing the response variable) is important in stating the conclusion, context displayed in stating the conclusion is considered in scoring component 2 of Section 1.

2022 Form I

- The response need not make an explicit decision about the null hypothesis (reject H_o or fail to reject H_o).

- If the conclusion and justification are consistent with an incorrect *p*-value (or an incorrect value of the test statistic, or an incorrect confidence interval, or an incorrect critical value), the response may satisfy component 1 and component 2.

- If the statement of the conclusion is incorrect AND the justification is inconsistent with the stated conclusion, then neither component 1 nor component 2 is satisfied. For example, the response correctly indicates that the *p*-value is less than the significance level and makes an incorrect conclusion that the null hypothesis should be not be rejected and there is not convincing statistical evidence that mean weight loss for overweight large-breed dogs fed Diet J is greater than the mean weight loss for overweight large-breed dogs fed Diet W. The statement about the *p*-value is correct, but the conclusion is incorrect and the conclusion is not justified by the statement about the *p*-value.

- To satisfy component 2, the conclusion must be consistent with the stated alternative hypothesis. If the response omits hypotheses, assume the correct alternative hypothesis, $H_a: \mu_J > \mu_W$, was provided when scoring component 2.

- If the response provides the incorrect comparison between the stated p-value and the level of significance, but the conclusion is consistent with the given comparison and the alternative hypothesis, then component 2 is satisfied.

- If the conclusion includes a definitive statement (e.g., "this proves that the claim of a difference is not true …", or "there is no evidence…"), then component 2 is not satisfied.

- If the response includes a statement that is equivalent to accepting the null hypothesis (e.g., "we conclude that the mean weight loss of all overweight large-breed dogs is the same for the two diets"), then component 2 is not satisfied.

- If the response includes an incorrect interpretation of the *p*-value (e.g., "this is the probability that the null hypothesis is true", then the score for Section 3 is lowered one level (that is, from E to P or from P to I).

- In Section 3, the conclusion may reference whether or not the diets *tend* to have different weight loss, or that "we cannot conclude that the two diets have different *tendencies.*" This can be read as "central tendency," and therefore can be interpreted as the "*population mean.*"

- If the response incorrectly identified a two-sample *z*-test for a difference in population means, then section 3 should be scored using the reported *p*-value (or comparison of the *z*-value to a critical value obtained from the standard normal distribution).

Confidence Interval Approach*:*

- Component 2 should be scored according to the rubric and component 1 should be scored with regard to a comparison of zero to the appropriate end of the reported confidence interval.

- If no alternative hypothesis is specified in the response, then assume the correct alternative hypothesis is provided when scoring component 2.

- If an incorrect two-sided alternative hypothesis is specified, then component 2 is satisfied if the justification is based on whether zero is included in the confidence interval.

- If the response includes an incorrect interpretation of the confidence interval, then the score for Section 3 is lowered from E to P or from P to I.

2022 Form I

Matched-Pairs Approach*:*

- Components should be scored according to the rubric using the reported *p*-value, or by comparing the *t*-statistic to an appropriate critical value, or comparing zero to the appropriate end of the reported confidence interval.

Rejection Region Approach*:*

- A response that uses a comparison of the value of the test statistic to a critical value in lieu of reporting and comparing the *p*-value to a significance level may satisfy component 1. The one-sided test critical values for a significance level of 0.05 for commonly used degrees of freedom (df) are shown in the table. The critical values for a two-sided test are also shown.

df	Critical Value (one-sided test)	Critical Value (two-sided test)
30	1.697	2.042
28.75	1.700	2.046
15	1.753	2.131

45. Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

A survey conducted by a national research center in 2018 asked a random sample of people over age 12 years old in the United States whether it is very important to them personally to have a lot of money. Responses are summarized based on age in the following table of counts.

	Adults	**Teenagers**
Very important	281	86
Not very important	180	168
Total	461	254

(a) Construct a graph to compare the responses of adults and teenagers.

(b) A hypothesis test will be conducted to investigate whether the proportion of all United States adults who would respond that it is very important to have a lot of money is different from the proportion of all United States teenagers who would respond that it is very important to have a lot of money. Identify the appropriate test and state the hypotheses.

(c) The p-value of the hypothesis test was found to be less than 0.0001. Based on the p-value, what is the appropriate conclusion for the test? Justify your answer.

Part A, Part B, Part C

Select a point value to view scoring criteria, solutions, and/or examples and to score the response

2022 Form I

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4.
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Scoring

Parts (a), (b), and (c) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

0	1	2	3	4 ✓

All three parts essentially correct.

- ☐ Part (a) essentially correct
- ☐ Part (a) partially correct
- ☐ Part (a) incorrect
- ☐ Part (b) essentially correct
- ☐ Part (b) partially correct
- ☐ Part (b) incorrect
- ☐ Part (c) essentially correct
- ☐ Part (c) partially correct
- ☐ Part (c) incorrect

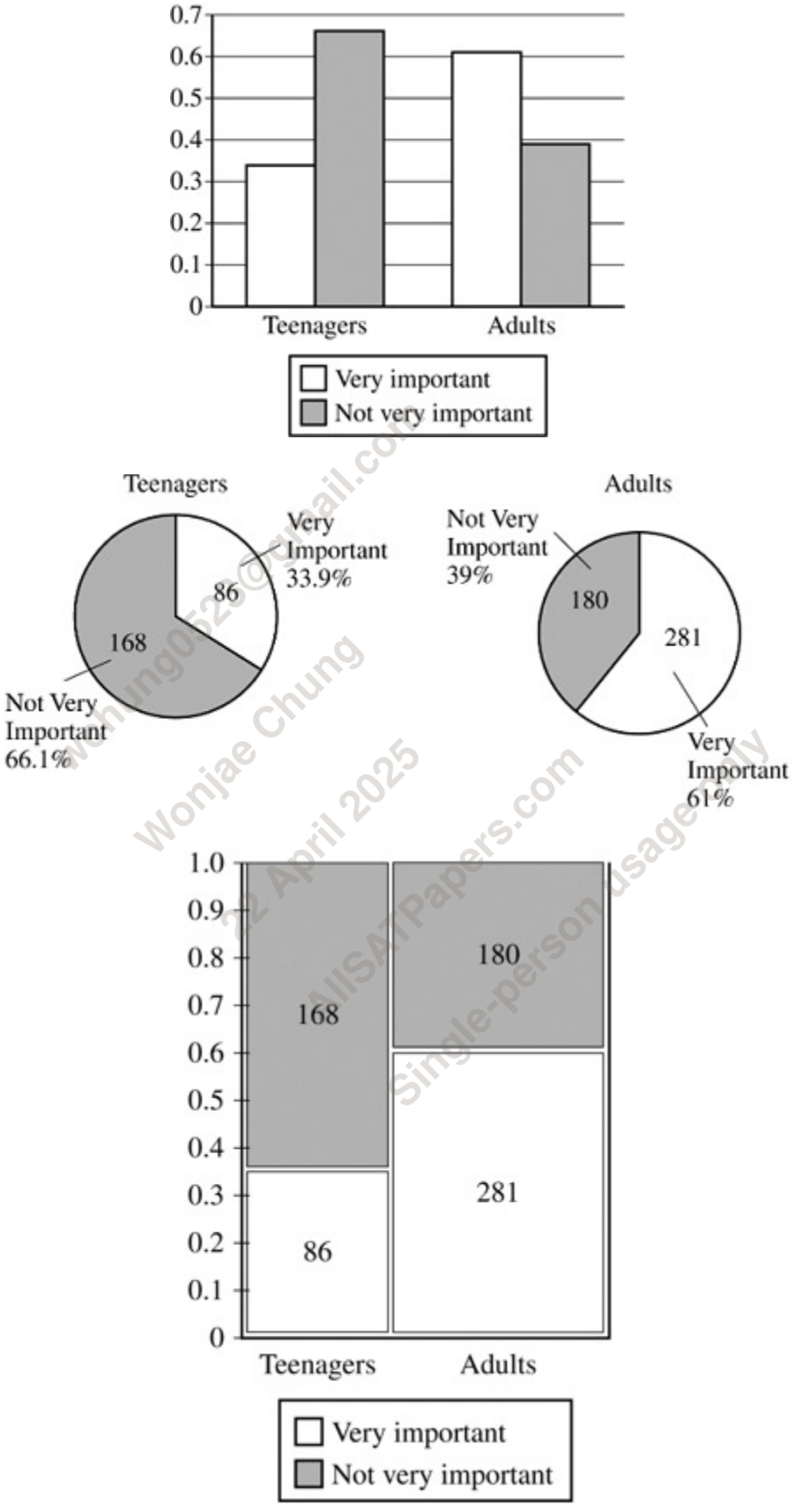
Solution

Part (a):

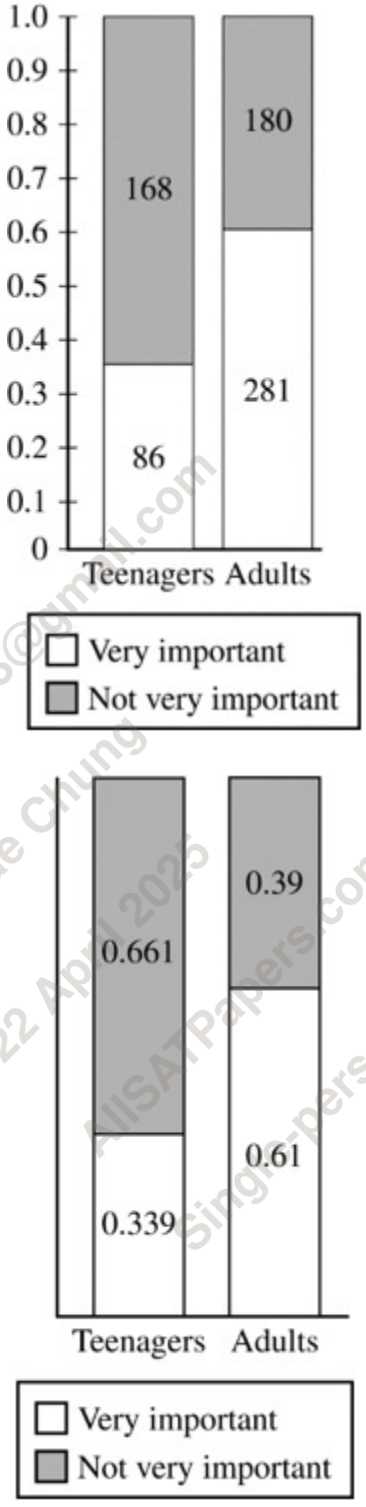
The sample proportion of adults who responded that having a lot of money is very important is $\frac{281}{461} \approx 0.610$. The sample proportion of teenagers who responded that having a lot of money is very important is $\frac{86}{254} \approx 0.339$.

Examples of reasonable graphs include:

2022 Form I



2022 Form I



Scoring

Part (a) is scored as follows:

Essentially correct (E) if the response constructs a correct graph that allows for the comparison of the responses of adults and teenagers that satisfies the following four components:

2022 Form I

Reflects the correct relative proportions or relative percents

Uses an appropriate scale to indicate counts or relative proportions (percents)

Uses labels, or includes shading and a legend, that clearly distinguish between very important and not very important responses

Uses labels that clearly indicate responses for adults and responses for teenagers

Partially correct (P) if the response constructs a graph that allows for the comparison of the responses of adults and teenagers that either satisfies component 1 and only one or two of the other three components required for E

OR

if the response satisfies only components 2, 3, and 4.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response that conditions on the response variable outcomes (very important and not very important) may receive credit for components 2, 3, and 4, but not component 1.
- Numerical values of the relative proportions are not necessary to satisfy component 1.
- For side-by-side pie charts, both components 2 and 3 are satisfied by labeling of the sections of each pie chart.
- A response that attempts to represent all data for both adults and teenagers on a single pie chart, or a single bar, should be scored I.

Solution

Part (b):

The appropriate test is a two-sample *z*-test for a difference between two population proportions. The corresponding hypotheses are

$H_0 : p_{adults} = p_{teenagers}$
$H_a : p_{adults} \neq p_{teenagers}$

where $p_{teenagers}$ is the proportion of all teenagers in the United States who would respond that it is very important to have a lot of money and p_{adults} is the proportion of all adults in the United States who would respond that it is very important to have a lot of money.

Scoring

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:

States the correct null hypothesis of equal proportions AND the correct two-sided alternative hypothesis

Provides sufficient context by including references to

- The two populations (adults and teenagers)
- Population parameters: the proportion of all adults (proportion of all teenagers) who would respond that it is very important to have a lot of money
- The response variable outcome (very important or not very important)

The two-sample z-test for a difference between two population proportions is identified by name
Partially correct (P) if the response does not meet the criteria for E but satisfies either component 1 and/or component 3.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- The hypotheses may be stated in words, or they may be stated using parameters.
- Two-proportion z-test is sufficient for component 3, but two-sample z-test, without an indication of proportion in the name of the test, is not sufficient.
- Notation must clearly distinguish between adults and teenagers. This could be done, for example, by using p_a and p_t to distinguish between the two proportions.
- To satisfy component 1, the hypotheses must be stated in terms of proportions. Any parameter notation other than p or π must be clearly defined as the population proportion.
- Parameters that are written in the case of "Not very important" are acceptable as long as the parameters are clearly defined. For example: The hypotheses would be $\begin{array}{l} \mathrm{H_0}: p_{adults} = p_{teenagers} \\ \mathrm{H_a}: p_{adults} \neq p_{teenagers} \end{array}$ where $p_{teenagers}$ is the population proportion of all teenagers in the United States who would respond that it is not very important to have a lot of money and p_{adults} is the population proportion of all adults in the United States who would respond that it is not very important to have a lot of money.
- A chi-square test for homogeneity or independence may satisfy component 3, if the alternative hypothesis is two-sided.
- If the chi-square test of independence is identified, the response may satisfy component 1 with hypotheses written as associations, such as $\mathrm{H_0}$: There is no association between age group and whether money is very important, $\mathrm{H_a}$: There is an association between age group and whether money is very important.
- A response based on the chi-square test of independence provides sufficient context by including references to the two populations (all US adults and teenagers), the association, and the response variable outcome (very important or not very important)
- Conditions (e.g. normality, random selection) are extraneous.
- Only the response to part (b) may be used for scoring part (b). The response to part (c) may not be used to score part (b).

Solution

Part (c):

Because $p < 0.0001$ is smaller than any reasonable significance level, there is significant evidence to show that the

proportion of all adults who would respond that it is very important to have a lot of money is different from the proportion of all teenagers who would respond that it is very important to have a lot of money.

<u>Scoring</u>

Part (c) is scored as follows:

Essentially correct (E) if the response includes the following two components:

Provides justification of the conclusion based on a correct comparison of the given p-value with a reasonable significance level (e.g., 0.01, 0.05, 0.10, etc.) or states the p-value is very small

Provides a correct conclusion, in context, that is consistent with the alternative hypothesis stated in part (b)

Partially correct (P) if the response satisfies only one of the two components

OR

Component 1 is not satisfied, but component 2 states a correct conclusion but it is not in context. For example, the response simply indicates that there is sufficient evidence to reject the null hypothesis.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- If the response to part (b) expresses the alternative hypothesis with respect to the proportions who would respond "Not very important," the conclusion should be expressed as follows:

  "Because $p < 0.0001$ is smaller than any reasonable significance level, there is significant evidence to show that the proportion of all adults who would respond that it is <u>not</u> very important to have a lot of money is different from the proportion of all teenagers who would respond that it is <u>not</u> very important to have a lot of money."

- If the statement of the conclusion is incorrect AND the justification is inconsistent with the stated conclusion, then neither component 1 nor component 2 is satisfied. For example, the response correctly indicates that the p-value is less than a certain significance level (or is small) and makes an incorrect conclusion that the null hypothesis should be not be rejected and there is not convincing statistical evidence that proportion of all adults who would respond that is important to have a lot of money is different than the proportion of teenagers who would respond that it is important to have a lot of money. The statement about the p-value is correct, but the conclusion is incorrect and the conclusion is not justified by the statement about the p-value.

- To satisfy component 2, the conclusion must be consistent with the stated alternative hypothesis. If the response omits hypotheses, assume the correct alternative hypothesis, $\mathrm{H_a}: p_{adults} \neq p_{teenagers}$, was provided when scoring component 2.

- If the conclusion includes a definitive statement (e.g., "this proves there is a difference…", or "there is no evidence…"), then component 2 is not satisfied.

- If the response includes a statement that is equivalent to accepting the null hypothesis (e.g., "we conclude that the population proportions are the same for adults and teenagers"), then component 2 is not satisfied.

- If the response includes an incorrect interpretation of the p-value, then the score for part (c) is lowered one level (that is, from E to P or from P to I).

2022 Form I

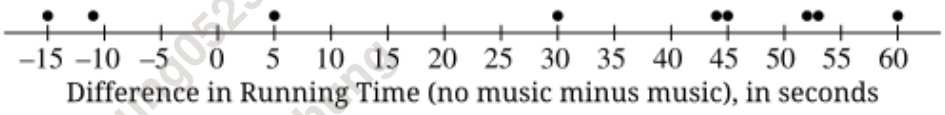
- If the response provides the incorrect comparison between the stated *p*-value and the level of significance, but the conclusion is consistent with the given comparison and the alternative hypothesis, then component 2 is satisfied.
- At a minimum, context needs to refer to the proportion of adults and teenagers. If the response refers to the chi-square test of independence in part (b), sufficient context can be satisfied by referring to the association between adults and teenagers.
- Only the response to part (c) may be used for scoring part (c). The response to part (b) may not be used to score part (c).

2022 Form I

46. Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Nine experienced runners participated in a well-designed controlled experiment to investigate whether listening to music while running will allow a runner to run faster. On each of two consecutive Saturdays, the runners ran on a track until they completed 3.1 miles. Researchers recorded the time, in seconds, for each runner to complete 3.1 miles. On one day the runners listened to music while running, and on the other day the runners did not listen to music. For each runner, a coin was flipped to determine on which day, the first or second, music was used.

The following dotplot shows the differences in running times (no music minus music) for the 9 runners.



(a) For the investigation, what does the positive difference of 60 indicate?

If listening to music has no effect on running time, then running time is equally likely to be greater for either condition, music or no music. Under that assumption, the probability that the difference (no music minus music) will be positive for any particular runner is $p = 0.5$. Let X represent the number of runners out of 9 runners who have a positive difference (no music minus music) in running time.

(b) Assume that the differences in running times are independent of one another. From the sample of 9 runners, 7 have a positive difference (no music minus music).

(i) If listening to music has no effect on running time, what is the probability that among 9 runners, 7 or more runners will have a positive difference in running times? Show your work.

(ii) Based on your answer in part (b-i), is it reasonable to believe that runners tend to have longer running times while not listening to music? Explain your answer without performing an inference procedure.

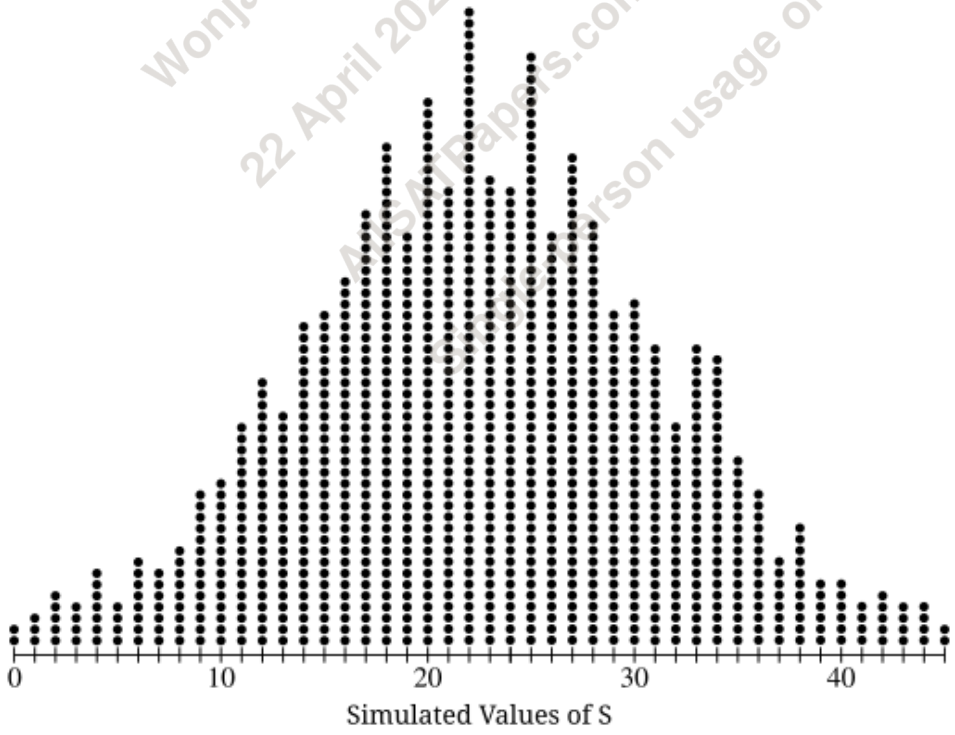
In part (b-ii), only the signs of the differences (positive or negative), not the magnitudes, are used. An alternative method that takes into account both the sign and the magnitude of the differences in running times is based on the sum, S, of ranks corresponding to positive differences. For the experiment described above, S is computed by first ordering the absolute values of the nine differences in times from smallest to largest. Next, ranks are computed for the absolute values of the differences with rank 1 corresponding to the smallest absolute value, rank 2 corresponding to the second smallest absolute value, up to rank 9 corresponding to the largest absolute value. Finally, the value of S is computed by summing only the values of the ranks corresponding to positive

differences; the values of the ranks corresponding to negative differences are not included in the sum. The following table displays the differences and the ranks of the absolute values of the differences for the nine runners who participated in the experiment, where rank 1 indicates the smallest absolute value and rank 9 indicates the largest absolute value.

Differences	5	-11	-15	30	44	45	52	53	60
Ranks of absolute values of the differences	1	2	3	4	5	6	7	8	9

(c) Determine the value of S.

A simulation was carried out to investigate the sampling distribution of S. The value of S was calculated for each of $1{,}000$ trials of the simulation, where each trial was simulated under the assumption that listening to music has no effect on running time. The results are shown in the dotplot.



(d) Based on the simulation results and your response to part (c), is it reasonable to conclude that not listening to music tends to result in longer running times? Justify your conclusion.

Part A, Part B, Part C, and Part D

2022 Form I

Select a point value to view scoring criteria, solutions, and/or examples and to score the response

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4.
- The model solution represents an ideal response to each part of the question and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Scoring

Parts (a), (b), (c), and (d) are each scored as essentially correct (E), partially correct (P), or incorrect (I).

0	1	2	3	4 ✓

Parts (a) through (d) sum to 4 points.

OR

Parts (a) through (d) sum to $3\frac{1}{2}$ points AND a holistic approach is used to decide to score up.

- ☐ Part (a) essentially correct
- ☐ Part (a) partially correct
- ☐ Part (a) incorrect
- ☐ Part (b) essentially correct
- ☐ Part (b) partially correct
- ☐ Part (b) incorrect
- ☐ Part (c) essentially correct
- ☐ Part (c) partially correct
- ☐ Part (c) incorrect
- ☐ Part (d) essentially correct
- ☐ Part (d) partially correct
- ☐ Part (d) incorrect

Solution

Part (a):

A positive difference occurs when the running time for a particular runner is greater when the runner is not listening to music than when the runner is listening to music. The difference of positive 60 means that it took the runner 60 seconds longer to run 3.1 miles when not listening to music than when listening to music

Scoring

2022 Form I

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

Identifies 60 seconds as a difference between a pair of observations for a single runner

Explains that the running time without music is greater than with music

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not meet the criteria for E or P.
Additional Notes:

- A response that gives incorrect units (e.g., 60 minutes, 60 hours) does not satisfy component 1.
- Equivalents to 60 seconds (1 minute, etc.) are acceptable substitutes when scoring component 1.
- Extraneous statements about causality should be ignored in scoring part (a).

Solution

Part (b):

(i) Let the random variable X represent the number of the 9 runners for which the difference in running times is positive. Under the assumption that listening to music has no effect on running time, X would have a binomial distribution with parameters $n = 9$ and $p = 0.5$. Thus,

$$P(X \geq 7) = \binom{9}{7}(0.5)^7(1-0.5)^2 + \binom{9}{8}(0.5)^8(1-0.5)^1 + \binom{9}{9}(0.5)^9(1-0.5)^0 = \frac{46}{512} \approx 0.090.$$

(ii) The probability of 0.09 is large when compared to a significance level of, for example, 0.05, indicating that it would not be very unusual to see 7 or more positive differences if listening to music has no effect on running time. Therefore, at a 0.05 significance level, there is not convincing evidence to conclude that the running time when not listening to music tends to be greater than when listening to music.

Scoring

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following four components:

The response to part (b-i) provides the correct probability

The response to part (b-i) shows sufficient supporting work by satisfying both of the following:

- indicates the use of a binomial distribution with parameters $n = 9$ and $p = 0.5$

2022 Form I

- clearly identifies the event for which the probability is to be calculated, e.g., $X \geq 7$

The response to part (b-ii) correctly compares the probability calculated in part (b-i) to any valid level of significance, or comments on the "small" or "not small" size of the probability

The response states a correct conclusion based on the comparison from component 3

Partially correct (P) if the response satisfies only two or three of the four components.

Incorrect (I) if the response does not satisfy the criteria for E or P.

Additional Notes:

- For component 2, use of a binomial distribution with parameters $n = 9$ and $p = 0.5$ may be accomplished by one of the following:
    - Words or standard notation: Defining a random variable, say *X*, and either stating in words that *X* has a binomial distribution with $n = 9$ and $p = 0.5$ *OR* using common notation such as $X \sim B(9, 0.5)$.
    - Formula: Displaying a binomial probability formula with 9 inserted for *n* and 0.5 inserted for *p*.
    - Calculator function syntax: Using calculator function notation with *n* and *p* clearly labeled, e.g., $1 - \text{binomcdf}(n = 9, p = 0.5, \text{upper bound} = 6)$.
- For component 2, the event for which the probability is to be calculated may be indicated by one of the following:
    - Random variable: Defining a random variable, say *X*, and using a probability statement, such as $P(X \geq 7)$ or $1 - P(X \leq 6)$.
    - Equation: Displaying a sum of probabilities, such as $\binom{9}{7}(0.5)^7(1-0.5)^2 + \binom{9}{8}(0.5)^8(1-0.5)^1 + \binom{9}{9}(0.5)^9(1-0.5)^0$.
    - Graphical display: e.g., a well-labeled bar chart with bars for outcomes 7, 8, and 9 clearly indicated.
    - Calculator function syntax: Using calculator function notation with *n*, *p* and bounds clearly labeled, e.g., $1 - \text{binomcdf}(n = 9, p = 0.5, \text{upper bound} = 6)$.
- Component 2 is not satisfied if a normal approximation to the binomial distribution is used to compute the probability.
- Using calculator function notation, such as $1 - \text{binomcdf}(n = 9, p = 0.5, \text{upper bound} = 6)$, may satisfy both parts of component 2 if the parameters and bounds are clearly identified.
- A response to part (b-ii) that compares the value calculated in part (b-i) to α, without specifying a value for α, may satisfy component 3.
- A response to part (b-ii) that simply describes the probability calculated in part (b-i) as "statistically significant" or "not statistically significant" does not satisfy component 3.
- Component 4 is not satisfied if the response to part (b-ii) interprets the value computed in part (b-i) as a probability of a positive difference in running time.

2022 Form I

- Component 4 is not satisfied if the conclusion includes a definitive statement (e.g., "this proves that not listening to music results in longer running times" or "there is no evidence that not listening to music results in longer running times").
- Component 4 is not satisfied if the response includes a statement that is equivalent to accepting a null hypothesis (e.g., "we conclude that listening to music has no effect on running time").
- The quality of the response to part (b-ii) may be considered if holistic scoring is necessary.

Solution

Part (c):

Differences	5	−11	−15	30	44	45	52	53	60
Ranks of absolute values of the differences	1	2	3	4	5	6	7	8	9

Summing only the ranks (of absolute differences) which correspond to positive differences gives $S = 1 + 4 + 5 + 6 + 7 + 8 + 9 = 40$.

Scoring

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

Computes a correct value of *S*

Provides supporting work which either displays the formula for the sum of the correct ranks

OR

Indicates the ranks used in the sum by marking values in the table

Partially correct (P) if the response satisfies only one of the two components required for E.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Notes:

- A response that sums the values of the positive differences instead of the corresponding ranks may be scored P if the sum of 289 is given with supporting work displaying the formula for the sum of the positive differences OR with an indication that the positive differences were used by marking values in the table.
- A response that displays the formula for the sum of the correct ranks but does not correctly compute S will satisfy component 2 but not component 1.

A response that explicitly sums the ranks $4 + 5 + 6 + 7 + 8 + 9$ to yield $S = 39$ is scored P.

Solution

2022 Form I

Part (d):

Yes, based on the simulation analysis of sums of ranks of differences, it is reasonable to conclude that not listening to music tends to result in longer running times. The observed value of the test statistic *S* is 40, and the simulation analysis reveals that obtaining a test statistic value at least this large only occurred in 25 of the 1,000 simulated trials that were carried out under the assumption that music has no effect on running time. Because this indicates an unlikely event, $p - \text{value} = 0.025$, it is reasonable to conclude that not listening to music tends to result in longer running times.

Scoring

Part (d) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:

Determines the percentage (or number) of simulated test statistic values that are at least as large as 40 (or the value of *S* computed in the response to part (c))

Uses the numerical evidence to justify an appropriate conclusion about whether not listening to music tends to result in longer running times (slower running)

Partially correct (P) if the response satisfies only one of the two components required for E

OR

states an appropriate conclusion AND justifies it by noting the relative location of 40, or the value of *S* computed in the response to part (c), in the dotplot, but the response does not provide specific numerical evidence.

Incorrect (I) if the response does not justify the conclusion based on where the value of *S*, computed in the response to part (c), falls in the dotplot

OR

otherwise does not meet the criteria for E or P.

Additional Notes:

- Although the testing procedure based on *S* is not exclusively designed to detect a change in a mean, responses that use the numerical evidence to justify a conclusion about the difference in mean running times, i.e., listening to music reduces the mean running time, may satisfy component 2.
- A response that counts only the values equal to 40 or only the values greater than 40 does not satisfy component 1.
- A response that correctly computes the percentage but refers to this percentage as the probability that *S* is greater than 40 (instead of greater than or equal to 40) may satisfy component 1.
- If the response determines the percentage (or number) of values equal to or greater than 40 plus those equal to or less than 5, which is $25 + 25 = 50$ (or the two-sided *p*-value corresponding to the value of *S* computed in the response to part (c)), then component 2 can be satisfied if that numerical evidence is used to justify a two-sided conclusion OR a right-tailed conclusion.
- A response that uses a normal approximation to compute the probability that a simulated value of the test statistic is at least 40 (or the value *S* computed in part (c)) does not satisfy component 1 but the calculated probability

2022 Form I

may be used to satisfy component 2.

- Component 2 is not satisfied if the conclusion includes a definitive statement (e.g., “this proves that not listening to music results in longer running times” or “there is no evidence that not listening to music results in longer running times”).
- Component 2 is not satisfied if the response includes a statement that is equivalent to accepting a null hypothesis (e.g., “we conclude that listening to music has no effect on running time”).
- The quality of the response to part (d) may be considered if holistic scoring is necessary.